数据的真相

如何在数字时代做出明智决策

EVERYDATA

The Misinformation Hidden in the Little Data You Consume Every Day

[美] 约翰·H.约翰逊
[美] 迈克·格鲁克 ◎著
王 喆◎译

MIT科学家解读数据的本质

消费者及市场、营销、广告、公关从业人员相见恨晚的一本书

营销大师 赛斯·高汀 推荐

中信出版集团

版权信息

书名:数据的真相：如何在数字时代做出明智决策

作者:[美]约翰·H.约翰逊 [美]迈克·格鲁克

译者:王喆

ISBN:9787508689982

中信出版集团制作发行

前言

2010年春，美国国家橄榄球联盟（NFL）的老板们在提出诸多要求之外，还想将常规赛季的比赛场数额外增加两场。作为国家橄榄球球员协会的执行主任，我觉得这并不是一件好事。橄榄球赛的受伤率已经高达100%，因此多打比赛意味着更多伤病。而伤病率的上升会导致运动员职业生涯进一步缩短，现在运动员的运动生涯本就只有三年多一点。橄榄球运动员在赛季期间每周都会冒着很大风险投身于这项美国人民最为喜爱的运动中，我们协会对他们的安全、报酬以及保护措施极为关注。

职业橄榄球产业创造的价值高达数十亿美元，而且依然在不断发展壮大之中。但在2010—2011年间，该行业经历了数十年不遇的停摆，因此我需要以一个具有说服力的方式来证明赛季的长度必须维持在十六场比赛不变。

于是我拨通了约翰·约翰逊的电话。

约翰是经济学、统计学、数据学领域的专家，他能够把极为复杂的数据理论变得简单易懂，让人一目了然。简而言之，他应该能够算得上是我认识的人中最为聪明的一个。他能够仔细、全面地解释问题，不论你受教育水平如何，都能够听得懂。而且最重要一点，他这人还很风趣。

怀着和美国所有橄榄球迷一样的心愿，国家橄榄球球员协会想拯救即将到来的全国橄榄球联赛，因此我们把相关数据交给了约翰。约翰及其同事开发出一个模型，展示球员受伤的时间和频率。他们指出导致最严重伤害的动作（从而引出对球员安全造成巨大影响的“开球规则”）。

他们通过数据预测出平均每个球员职业生涯缩短的程度，估计出将全国橄榄球联赛赛季延长至十八场比赛可能导致的经济损失。他们同我们合作，将每场全国橄榄球联赛对周边社区产生的实际价值进行量化，帮助我们对可能发生的结束赛季停摆所产生的经济影响有一个更好的理解。

最终，约翰以其出色的口才向球员、媒体，尤其是美国国家橄榄球联盟的老板们（他们对“损失”这类词特别关心）解释了这些数据，我们得以做出既符合我们初衷，又能保障球员的健康和安全的决定，最终促成了后续的谈判，签订了一项历史性的协议，该协议为期十年，很好地保障了美国国家橄榄球联盟球员的权益。

本书很好地反映了约翰的思想。我对该书能够帮助你在日常生活中做出更好的决定十分有信心。通过阅读本书，你可以学到如何对每天接收到的数据进行理解、解读、思考。约翰和本书的合著者迈克·格鲁克通过列举数百个例子，直指问题本质，举重若轻地将复杂问题化繁为简（有时候甚至能让你忍俊不禁）。因此在阅读本书过程中，要是发现自己因为恍然大悟而频频点头，并开始质疑家里和单位中看到的每一项“事实”，请不要感到惊讶。

有一个不可否认的事实，即你每天生活中的数据正在迅速增加，并从四面八方向你涌来。因此了解如何利用数据，以及在什么情况下数据会被滥用十分重要。我亲身体验了数据的力量，但你不需要和一群世界上最富有的人一起死命加班，体会理解数据、管理数据的重要性。在你看电视、购物、工作、在餐厅吃饭的时候，这本书是你的不二之选。《数据的真相》见解深刻、文字富有魅力、内容引人入胜，在关键时刻填补了一项关键空白。

本书相当于橄榄球队主教练编写的比赛战略手册，能够助你在运用数据的赛场上变得更强大、更聪明、更有信心。

祝阅读愉快。

莫里斯·F. 史密斯

美国国家橄榄球球员协会

执行主任

序

一万个问题亟待解决

见鬼，明尼苏达州究竟发生了什么事？

2015年夏天，10天内有关方面宣布召回三种被沙门氏菌污染的食品。这三类问题食品的源头都在明尼苏达州，当地人吃坏了肚子。

涉事企业最终召回了约370万磅鸡肉制品，除此之外还有数量不明的黄鳍金枪鱼[注]。

但是，为什么是明尼苏达州？

是不是所有受到污染的食物都产自明尼苏达州？不是。那是不是都从那儿装船运输？不是。是不是明尼苏达州的居民更容易感染沙门氏菌导致的疾病？据我们所知，也并非如此。

据雅虎健康网的报道，此事件的内部联系十分简单：明尼苏达州和美国其他州相比，在诊断由食物引起的疾病方面水平更高。[注]

有些人可能听说有三起食品召回事件都发生在明尼苏达州，就认为那个地区的食品不安全。但事实并非如此。恰恰相反，由于明尼苏达州健康农业部工作出色，你的健康在那里可能比在美国其他地区得到更好的保障。

每天你都被媒体报道或其他来源的消息所包围，这些消息通常充满了隐藏信息——或误导信息。本书将帮助你辨别、解读这些信息，使你

成熟地汲取数据（“信息”或“事实”比较高端的说法）。

全书通篇将回答以下问题：

一则虚假新闻如何让股市一下子蒸发掉1360亿美元？（第六章）

福岛核灾难是否可以避免？（第八章）

4/5的儿科医生是如何看待婴儿食品的？（第七章）

如何知道哪个总统候选人的选票真正领先？（第五章）

吃烤奶酪三明治的人们真的更为“性福”吗？（第四章）

“小数据”不是一个常用词。的确，现在还不是。我们造出这个词，用来描述你每天碰到的数以吨计的数据。我们撰写了这本书，帮助你在理解小数据时，更得心应手、更灵活、更快捷。

因为哪怕你并没有接受过成为数据专家的相关训练，你依然需要解读数据。如果你是名律师，你需要看出证人是否在陈述证词的时候有所保留。如果你是名护士，你必须理解患者发热的症状是由疾病引起的，还是仅仅与疾病相关。如果你是一名首席执行官，你会想要预测下个季度收益为多少。如果你是一名家长，你会想要知道自己孩子的能力是否超过（或低于）平均水平。

这本书讲的是如何利用生活中的数据做出更好的决策。

纵观全书，我们会运用真实或假设的例子，重点指出经常被曲解的数据概念。在每一章的结尾，我们会列出5种让你运用所学知识的方式。你可以将本书从头读到尾，也可以直接挑你感兴趣的章节读。

本书开篇，记述了脸谱网上两个老友一段简单的留言对话。这段留言的时间为2013年秋天，留言双方为约翰（经济学家、统计学家）和迈

克（作家）。

“嗨，迈克——我想写本书，脑子里多少有点概念……要不要找个时间聊聊？”

约翰作为职业经济学家，有着同世界各地公司商量如何解读其数据的经验。基于这些经验，他和迈克分享了想要让数据概念变得通俗易懂的想法后，书的脉络渐渐成形。

“明年一整年我们可能会为这件事乐此不疲。”约翰写道。

事实果真如此。（准确说来，是整整两年。）

这条脸谱网留言下的回复一来二去积攒到了3288条。我们期间也发现我们生活中接触的数据比想象中更多。不论何处，只要稍加观察，就能发现越来越多数据被误读、曲解，或者干脆数据本身就是错误的例子。

于是就有了这本书，原本只是在脑中挥之不去的想法，变成了实实在在的文字。

希望你们能喜欢。

在切入正题之前，先呈上一个免责声明：约翰作为职业经济学家和统计学家，他的日常工作包括作为专家证人对数据进行缜密的分析，为《财富》杂志评选出的世界前100强公司、贸易集团、政府机构打涉及数百万美元的官司。(注)本书只对关键的数据概念做一个基本的概括性的论述，而非内容全面的教科书。因此如果你反对本书所推崇的让读者在阅读中能“会心一笑”这一特点，那你不用读下去了。有一个好消息，即本书兼具教育性和娱乐性，因此书中肯定有好玩的地方——如解释为什么你汽车的油箱不会完全空掉，为什么名人不会三人同时死亡，以及为什么喝过期牛奶没事。有了上面的免责声明在先，我们现在开始谈谈数

据吧。

---

1. “Osamu Corporation Voluntarily Recalls Frozen Yellow Fin Tuna Chunk Meat Due to Possible Health Risk,” U.S. Food and Drug Administration Recall Press Release, July 21, 2015, http://www.fda.gov/Safety/Recalls/ucm455622.htm; “Barber Foods Recalls Stuffed Chicken Products Due to Possible Salmonella Enteritidis Contamination,” United States Department of Agriculture News Release, July 12, 2015, http://www.fsis.usda.gov/wps/portal/fsis/topics/recalls- and- public- health- alerts/ recall- case-archive/archive/2015/ recall- 096- 2015- release; “Aspen Foods Recalls Frozen, Raw, Stuffed & Breaded Chicken Products Due to Possible Salmonella Enteritidis Contamination,” United States Department of Agriculture News Release, July 15, 2015, http://www.fsis.usda.gov/wps/portal/fsis/topics/ recalls- and- public- health- alerts/ recall- case-archive/archive/2015/ recall- 101- 2015- release.
2. Jenna Birch, “Why Recent Food Poisoning Outbreaks All Began in This One State,” Yahoo! Health, July 24, 2015, https://www.yahoo.com/health/why-do-so- many- foodborne-illness- outbreaks- begin- 124925339332.html. According to the article, Minnesota is better than other states because it is “incredibly thorough and quick” in terms of interviewing people who may be involved in cases, and because it does additional DNA analysis to identify the source of the illness. In other words, Minnesota is using data to solve the problem.
3. No confidential or client information has been used in the preparation of this book. This book represents only the views of the authors, not other economists or employees of Edgeworth Economics.

第一章 无处不在的数据：从大数据到小数据

从早晨睁开眼睛开始，你便被数据包围。根据“全球信息工业中心”（隶属于加州大学圣迭戈分校）[注]“多少信息”项目的研究结果，其实美国人平均每天接收约34千兆的数据。

34千兆可不是个小数目。1千兆——或平常会缩写成GB，相当于超过10亿字节（一个字节等同于一个字母或一个数字）。据BBC（英国广播公司）引用的说法，如果把34GB的数据打印出来，能够装满几十辆卡车。[注]这还仅仅只涵盖了你家中活动的数据源，如电视、收音机、电脑、电话等——这34GB的数据似乎并不包括我们周围存在的数据，也不包括我们工作中接收到的信息，如果把这些都算进去，那这个数据总量会是34GB的两三倍。[注]

我们来花上一分钟，想一想你在一天刚开始的一两个小时内可能碰到的“小数据”，并对其加以解读：

你睁开眼睛，看到了一天中的第一个数据——闹钟上幽幽亮着的数字。

接收数据：约9字节。

你拿起手机。哪怕在起床前，浏览十几封电子邮件，看几条信息，看一下交通情况，读几条重大新闻也并非难事。

接收数据：约2.1M（1M＝约100万字节）。

你走进浴室，站上体重秤，秤上的读数表明你昨晚不该吃比萨。

接收数据：约3字节。

你可以从牙刷上接收到数据吗？当然可以，只要是那种每30秒震动一次提醒你换个刷牙位置的电动牙刷就行。震动是数据的另一种表现形式。

接收数据：约60字节。

叮！你手机上的日历铃声提醒你要开一次客户会议。

接收数据：约43字节。

你服用多元维生素吗？吃不吃鱼油？你得仔细阅读标签内容，从而理解上面的所有数据。药品标签上的数据都非常重要。

接收数据：2M。

女儿要你在她的数学试卷上签字。你如何知道自己是否看到了所有数据？你女儿有没有把其他试卷藏起来？试卷上的分数又代表了什么？——代表了她的能力，或者仅仅与她平时上课出勤有关，还是有其他的打分标准？

接收数据：约46KB（1KB＝约1000个字节）。

你早晨是通过什么方式看新闻的？俄勒冈大学的研究表明，同阅读网络新闻相比，人们在阅读报纸时，能记得更多信息[注]。浏览新闻标题的时候，你看到以下信息：

最近投票显示有76%的美国人对国会工作不满。

波士顿棒球队红袜队排名第一——15场连胜，比第二名整整多出7

场。

天气预报显示最高气温为70华氏度（21℃），最低气温为58华氏度（14℃）。

研究表明每天喝一杯红酒能降低心脏疾病的发病率。

从政府预算可以看出美联储将再次下调利率。

接收数据：约3.1M。

你想知道一旦利率下调，重做购房贷款需要花费多少金钱与时间吗？你在智能手机的一个App（应用程序）中加上一条记录（该记录自动与计算机同步），并以电子邮件的形式发送给你的配偶。

接收数据：约2.2M。

该上班了。你一坐上车，马上就面对一个满是数据的仪表盘。机油灯是否亮着？发动机温度如何？胎压如何？有些数据会通过警示灯是否亮起显示，有些会通过刻度盘显示，有些则显示在电子屏上。

接收数据：约63字节。

你车子的油箱里还剩1/4箱燃油。这究竟说明了什么？为什么哪怕仪表盘上显示油箱已空，车子依然能够继续行驶？（答案见第六章）

接收数据：约26字节。

你到了加油站，公告牌上标示的汽油价格比路对面加油站每加仑便宜4美分，但油价便宜的加油站只收现金。每加仑4美分的差价是否足以影响你的加油选择？

接收数据：约2.0M。

你上班是坐公共交通工具还是自驾？如果你在华盛顿的话，可能会走495号州际公路，并且缴纳一笔根据行驶距离收取的“动态”通行费[注]。（换句话说，行驶过程中数据被收集，接着定价。）但是过路费的定价是否严格按照供需关系？或只是利用隐藏数据收取高额费用？

接收数据：约44.0M。

你到星巴克买咖啡。买16盎司的大杯还是20盎司的超大杯呢？你想了下价格差（还有卡路里）。点完咖啡，你看了一下善品糖、Sweet’N Low（一个甜品品牌）、怡口糖的包装，标签上都标明了营养和成分。（真的糖和蜂蜜没有营养成分表，因此比较难和这些竞争产品做比较。）

接收数据：约10.1KB。

你最终到了办公室，该工作了。看新闻、查电子邮件、审阅表格、协商合同、看销售预测、打电话、开会……所有的这一切都是数据，整天不断。

但就算是以上这些例子都没能真正解释数据是如何影响到我们生活的。因为上面的这些例子夹杂着去哪里加油、影响你孩子数学成绩有哪些因素之类的事，小数据能够影响到你的心情，影响到你能赚多少钱，甚至能影响到你的寿命。比如：

你应该在哪里买房或租房？你应该把更多注意力放在学校排名、上下班所花的时间、犯罪数据、房屋转手价值还是其他数据上？

如果你桌上有一份非常好的工作邀请函，你正犹豫着要不要接受。你应该运用什么数据来证明自己应得的薪资待遇水平？你把这些数据递交给那个可能成为你新老板的人，她会如何看待这些数据？

你该如何判断应该找什么样的人谈恋爱，或者有可能的话同其步入

婚姻殿堂？你会使用婚恋网站吗？在那些网站上，你的数据将经过过滤、解读，从而定位出一个与你相匹配的对象。什么数据对你来说最为重要？你父母的身高？对方毕业于哪所大学？还是对方能赚多少钱？

如果医生告诉你，你的血糖值或胆固醇水平超标，你会怎样？你体检得出的数值，有没有一个特定的临界值，超过以后就证明是危险的？上升的指数是源于基因、饮食，还是其他综合因素？医生可能会让你每天验血，并且记录饮食（从而为你和医生收集更多数据来做分析）。

以上这些情况你会如何处理？

“我们拥有的数据量很丰富，”《时代》杂志指出，“但数据给我们带来的回报正在迅速减少，因为一旦过了某个点，你掌握的信息越多，就越难理解其所包含的意义。”[注]

听起来是不是很耳熟？你并不是唯一这么想的。

---

1. Roger E. Bohn and James E. Short, “How Much Information? 2009 Report on American Consumers” (San Diego, University of California: Global Infor mation Industry Center, December 2009), 1–36, http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf.

2. The BBC cited a professor who estimated one GB as the equivalent of a pickup truck filled with printed pages. “Britons Growing ‘Digitally Obese, ” BBC News website, December 9, 2004, http://news.bbc.co.uk/2/hi/technology/4079417.stm.

3. That said, as the UC San Diego study explains, some of this data consumption happens simultaneously. For example, you might have the radio on as you work on your laptop.

4. Arthur D. Santana, Randall Livingstone, and Yoon Cho, “Medium Matters:Newsreaders’ Recall and Engagement with Online and Print Newspapers,”Print in the Mix website, accessed August 22, 2015, http://printinthemix.com/Research/Show/90.

5. As the organization that operates the Express Lanes explains it, “Sensors alongside the road monitor traffic levels and speed, and toll prices adjust to maitain free- flowing conditions in the Lanes— even during peak times— to provide value to customers.” “Using the Express Lanes,” Express Lanes website, accessed September 1, 2015, https://www.expresslanes.com/faqs.

6. Lev Grossman, “The Old Answer to Humanity’s Newest Problem: Data,”Time website,

June 25, 2015, http://time.com/3935273/ how- art- solves- the- data- problem/?iid=toc_062515.

“小数据”

你也许听说过大数据。所谓大数据，指的是那种量特别大，如果没有专业的软件和足够的计算机容量基本无法处理的数据[注]。比如联合包裹运输服务公司（UPS）每辆车收集200个数据点（该公司有大约10万辆车），利用这些数据节省空闲时间和燃料。彭博社有一篇文章写道：“每个司机每天少开1英里路程能够为公司节省折合高达50万美元的燃料费、车辆维护费以及时间。”[注]

世界各国利用大数据解决大问题。荷兰利用大数据管理水源。[注]食物运输公司利用大数据满足客户深夜进餐的需求。[注]IBM（国际机器公司）利用以PB计的数据确定可能发生的食品污染情况。（1PB＝1000000GB）[注]

大数据就像“性感美女”，能上头条。根据《福布斯》杂志一篇文章的说法，与大数据相关的工作需求每年以80%的速度增长[注]。《华尔街日报》引用与数据相关的项目的不断增长也说明了，商学院的学生“对大数据求知若渴”。[注]

但如你所见，正是那些小数据——那每天环绕你周围的以字节记的数据——对你每天的健康、开销、工作、人际关系等诸多方面产生了巨大影响。从食品标签到天气预报，从银行账户到医疗卫生，小数据无所不在。

不幸的是哪怕这些数据就摆在面前，人们依然对其心存疑窦。

《波士顿环球报》上的一篇文章指出：“事实并不总能改变我们的想法，能改变我们想法的反而并非事实。”[注]该文章引用密歇根大学的

一项研究，指出那些受误导的人们会坚信自己的看法，其中有些人甚至在面对事实的时候反而更坚定了自己错误的想法。（很明显，很多人只是不愿承认自己错了而已。）

但是，有句话说得好，传言再多也不能成为数据。比如你所有邻居都说今年夏天是有史以来最热的，并不能说明这就是真的。

艾米丽·奥斯塔尔是布朗大学经济学副教授，著有《你理应得到更好的：为什么传统妊娠智慧是错的以及什么才是你应该知晓的》一书，当我们问她人们如何在每天的生活中解读数据时，她说：“我认为我们所有人面对的最大问题是过分解读传闻。”

“人们十分乐于从个人故事或经历中学东西，”她继续说道，“你很难强迫自己无视身边的传言——或最低限度将其当作一个数据点来看——继而从其他数据总结出结论。”

传言可能让人印象深刻，可能听起来十分可信，但面对事实才是关键。

---

1. On its website, IBM defines big data as “data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low- latency.” “What Is Big Data Analytics?,” IBM website, accessed August 22, 2015, http://www-01.ibm.com/software/data/infosphere/hadoop/what-is- big- data- analytics.html.
2. Mary Schlangenstein, “UPS Crunches Data to Make Routes More Efficient, Save Gas,” Bloomberg Business, October 30, 2013, http://www.bloomberg.com/news/articles/2013-10-30/ ups- uses- big- data-to- make- routes- more- efficient save- gas.
3. Toby Wolpe, “Big Data Deluge: How Dutch Water Is Trying to Turn the Tide,”ZDnet website, October 1, 2013, http://www.zdnet.com/article/ big- data- deluge- how- dutch- water-is- trying-to- turn- the- tide/.
4. Alice Truong, “How Naturebox Uses Big Data to Stock Your Snack Pantry,”Fast Company website, June 4, 2014, http://www.fastcompany.com/3031078/fast- feed/ how- naturebox- uses- big- data-to- stock- your- snack- pantry.
5. “IBM Research Breakthrough Helps Public Health Officials Improve Food Safety,” IBM

website, July 3, 2014, accessed July 29, 2015, https://www-03.ibm.com/press/us/en/pressrelease/44295.wss.

6. Louis Columbus, “Where Big Data Jobs Will Be in 2015,” Forbes website,December 29, 2014, http://www.forbes.com/sites/louiscolumbus/2014/12/29/where- big- data-jobs- will-be-in-2015/.

7. Lindsay Gellman, “Big Data Gets Master Treatment at B-Schools,” Wall Street Journal website, November 5, 2014, http://www.wsj.com/articles/ big- data- gets- master- treatment-at-b-schools- 1415226291.

8. Joe Keohane, “How Facts Backfire,” Boston.com website, July 11, 2010, http:/www.boston.com/bostonglobe/ideas/articles/2010/07/11/how_facts_backfire/

小情境

这里还有一件你必须留心的事——在日常生活中，你也许会根据情境看数据，或将数据和你身边的其他数据做比较。有时候情境和其他数据是有用的，但有时候却会误导人。

试想：

在加利福尼亚州赫莫萨比奇市，消防队的平均反应时间约为5分钟多一点。[注]这个反应时间算不算快呢？为了解读这个数据，你可能要将其同这个城市过去的消防队反应时间做比较，或者在类似的城市间做比较，或者和其他数据做比较。

位于休斯敦的乔治·布什国际机场领导层接到了有关乘客行李到达时间过长的投诉。因此他们把取行李处转移到了离出站口很远的地方。旅客于是把时间都花在了走路上——而不是等行李上，投诉量减少了，因为人们觉得拿行李的时间缩短了。

媒介不同，我们接收、解读数据的方式也不同。比如你是否觉得《华盛顿邮报》网络版因为可以随时更新，所以比印刷版更为准确？你是否喜欢读纸质杂志，因为可以把有趣的文章裁剪下来？精装版的书是否比平装版的书看起来更为权威，尽管它们都是印刷品？

赛斯·戈登在一篇博文中问道："在一家餐厅，最便宜的红酒价格为30美元……而同样这款30美元的红酒是隔壁一家餐厅里最贵的酒……这款酒在哪个餐厅里喝起来味道更好？"[注]情境至关重要。

当然，还有一些你甚至可能不知其存在的隐藏数据。例如，根据《华尔街日报》报道[注]，如果你几年前在一个潮湿的天气查看一个名

为“天气频道”的应用，你也许会发现一个潘婷Pro–V洗发水的广告（专门修护发丝分叉）。可能这个广告只是个巧合，但事实上这个广告是根据不同邮编，专门以女性为目标投放的。如果湿度较高（会引起头发分叉），这则广告就会出现；如果湿度较低，则会出现其他头发护理产品的广告。你能看到天气预报，你也能看到广告，不过除非你在“天气频道”工作，否则你也许无法看出这两者之间的联系。在这个由数据驱动的世界，你并不总能了解到底是什么在驱动数据。

---

1. Data Analysis Report Fire and Emergency Medical Services Hermosa Beach, California (Washington, D.C.: ICMA Center for Public Safety Management,August 2013), 1– 52, http://www.hermosabch.org/modules/showdocument.aspx?documentid=3314.
2. Seth Godin, “Compared to What: Marketing and Relativity,” Seth Godin Blog, September 23, 2013, http://sethgodin.typepad.com/seths_blog/2013/09/marketing- and- relativity.html.
3. Katherine Rosman, “Weather Channel Now Also Forecasts What You’ll Buy,”Wall Street Journal website, August 14, 2013, http://www.wsj.com/articles/SB10001424127887323639704579012674092402660.

成熟的数据接收者

如果你从纽约来到新泽西，你可能记得赛姆斯服装店的电视广告，在广告里，赛·赛姆斯对观众说："成熟的客户是我们最好的客户。"（有趣之事：赛把他的姓按照店名改成了"赛姆斯"。注

赛说得没错。成熟的数据接收者远胜众人。

作为一个见多识广的数据接收者，你的任务便是不断问问题，从而理解自己所得到的数据是如何影响自己生活的。你需要问自己：

政党候选人在其电视广告中，有哪些东西没有说出来？

记者是否使用了精确的数据样本——抑或她只分享了能够支持自己文章的数据？

销售预测是基于哪些数据得出的？

你的医生说你的病是由某些行为引起的——还是这些行为只是和你的疾病相关？

市场推广人员在产品包装上重点突出了什么，为什么突出这些数据？

为什么年度报告上，有些数据以饼状图的形式出现，而其他数据以柱状图出现？

有些时候，数据会向你兜售东西——究竟兜售的是产品，是服务，还是观点？在这些情况下，只需知道新闻播报员想要得到你的关注，广告商想要得到你的钱，政客想要得到你的选票就能帮助你成为一个更好

的数据接收者。但有时候并没有这一整套流程——有的仅仅只是数据，你需要了解这些数据以理解周围的世界。

总而言之，这便是本书的主要内容：帮助你发现生活中的所有“小数据”，告诉你如何解读这些小数据，并为你提供切实有效的技巧来避免常见的数据陷阱，使你能够成为一个成熟的数据接收者，并且在日常生活中做出更好的决策。

我们这就开始吧。

---

1. Sy Syms was born Seymore Merinsky, but his family changed their name to Merns when Sy's dad and brother opened a store with the same name. Sy opened a competing store, but when he lost a legal fight to call it "Sy Merns," he renamed it SYMS— and then changed his own name to match. "Sy Syms, Founder of SYMS Corp., Dies at Age 83," PR Newswire website, November 17, 2009, accessed July 29, 2015, http://www.prnewswire.com/ news- releases/sy-syms- founder-of- syms- corp- dies-at-age-83-70407382.html.

第二章 对“挑战者号”评估结果的异议：抽样如何影响结果

“就在那残酷的瞬间，我们雀跃的心情一下子变成了恐惧；我们呆在原地看着眼前的一切，试着弄清楚到底发生了什么事。”[注]

1986年1月31日，罗纳德·里根总统站在约翰逊航天中心外，对几天前因“挑战者号”航天飞机在半空中解体而丧生的7名宇航员的家人、朋友、同事发表讲话。

之后数月，专家们花了数不清的时间采访主要目击者，审核相关证据，记录调查结果。

最终，问题的焦点落在了数据上。

负责航天飞机发射的人员犯了一个典型的错误，他们把注意力放在了错误的数据上。于是7个国家英雄“挣脱了大地粗暴的束缚”。[注]

---

1. Ronald Reagan, “Challenger Memorial Speech,” January 31, 1986, Johnson Space Center, Houston, Texas, YouTube, accessed April 25, 2015, https://www.youtube.com/watch?v=PhI9OQp6ADg.
2. President Reagan used this phrase— originally penned by British aviator John Gillespie Magee— in a speech to the nation on the day of the accident,at 5 p.m. EST on January 28, 1986, YouTube, https://www.youtube.com/watch?v=qoQlkFryriQ.

1986年1月28日

有一本记录此次事件的书是这么描写当时情况的："1986年1月28日上午，'挑战者号'航天飞机驶向卡纳维拉尔角空军站上方冰冷的蓝天，执行代号为51–L的任务。在兴高采烈的观众和屏息凝神的飞行控制员看来，这次发射一切正常。但是，升空后不到73秒，飞机的外挂燃料箱破裂，里面的液体燃料随之爆炸，'挑战者号'在空中解体。"[注]

《总统委员会关于"挑战者号"航天飞机事故的报告》指出："具体来说，问题出在防止热燃料泄漏的密封装置破损上……"[注]调查人员很快将注意力集中到了密封装置的关键部分——固态火箭发动机两个部件（U型接头和柄脚）之间的O型橡胶环上。

"挑战者号"上的O型环富有弹性，具备在几毫秒内收缩、膨胀的能力。但O型环的弹性"和温度直接相关……O型环在暖和的环境下会膨胀，填满U型接头和柄脚之间的空隙，但在冰冷的环境下可能就无法做到这一点"。[注]事实上，调查人员发现，处于压缩状态下的O型环在75华氏度（24℃）环境中的反应能力是在30华氏度（–1℃）环境中的5倍。

"挑战者号"发射时的气温为36华氏度（2℃）。[注]

《委员会报告》指出O型环"有可能"并没有像所需要的那样收缩、膨胀，从而出现了让燃料泄漏的空隙，摧毁了"挑战者号"[注]，这一点表明"哪怕美国最伟大的成就也可能因为一个像O型环那样平凡无奇的小物件而最终功亏一篑……"[注]

既然气温那么低，工程师们为何没有中止发射呢？

他们尝试过中止发射。鲍勃·隆德是莫顿聚硫橡胶公司工程部副总，该公司是固态火箭发动机的承包供应商。在那次灾难性发射的前夜，鲍勃·隆德和其他几个人提出建议，反对在寒冷的气候条件下发射“挑战者号”（该建议随后被撤销）。[注]

“我们担心气温会比去年一月份还要低，去年一月份的温度是50华氏度（10℃）还是53华氏度（12℃），我们发现……O型环有一定程度的磨损……而这并不是我们最担心的事，就像我们之前说过的，如果气温低于当时的51华氏度（11℃）或者53华氏度，不管具体几度，我们不知道事情能顺利开展到哪一步。我们担心的就是这种未知状态。”[注]

换句话说——他们手头的数据不够。没人知道在比之前任何一次发射气温低上15华氏度的环境中，O型环会出现什么问题。[注]

然而，缺少低于53华氏度的数据只是问题之一。

---

1. Andrew J. Dunar and Stephen P. Waring, Power to Explore— History of Marshall Space Flight Center 1960– 1990 (Washington, D.C.: National Aeronautics and Space Administration, NASA History Office, Office of Policy and Plans 1990), 339.

2. Report of the Presidential Commission on the Space Shuttle Challenger Accident (aka the Rogers Commission Report) (1986), chapter IV.

3. Rogers Commission Report, chapter IV.

4. According to the Rogers Commission Report, chapter IV, the temperature near the joint that failed was estimated to be 28 degrees Fahrenheit, +/- 5 degrees.The temperature on the opposite side of the booster— which faced the sun— was estimated to be 50 degrees Fahrenheit.

5. There were at least three potential issues with low temperatures. The first is O-ring resiliency. The second, as explained by the Rogers Commission Report (chapter IV), was “the potential for ice in the joints.” Finally, as the report noted,“O-ring hardness is also a function of temperature and may have been another factor in joint performance.”

6. Kevin Smokler, “The Day That Gen X Grew Up,” Baltimore Sun, February 3,2006.

7. According to the Rogers Commission, “the Thiokol Management reversed its position and recommended the launch of 51-L, at the urging of Marshall and contrary to the views of its

engineers in order to accommodate a major customer” (Rogers Commission Report, chapter V).

8. Bob Lund, quoted in the Rogers Commission Report, chapter V.

9. Various tests were actually conducted as low as 30 degrees. However, these experiments were conducted on test devices, and did not always include the putty that was part of the sealing system.

理解样本选择

你在给手头的数据样本做统计分析的时候，会引出一个在统计学中称为“样本选择”的问题。分析的数据数量比数据总量少并不一定会产生问题，但可能导致得出错误的结论，这一切取决于你要回答的问题是什么。在“挑战者号”这个例子中，数据被局限于失败这一块，但是失败的可能性恰恰正是人们所关心的问题，他们并没有对没有出现问题的样本进行研究。

负责该项目的团队清楚他们没有53华氏度以下的数据，于是决定在不考虑气温的前提下对O型环失灵的每一次情况展开研究。科学家和工程学家们根据数据得出的结论并没有错。但是，正如你所看到的，问题出在他们并没有根据自己想要回答的问题选择正确的数据进行研究。在当时情况下，他们应当对有关O型环运作状态的所有数据进行观察，而不应把研究对象局限于O型环产生失灵迹象的数据。

空难发生的前夜——工程师们试图说服莫顿聚硫橡胶公司和美国国家航空航天局（NASA）的管理人员取消发射——有人指出航天飞机在75华氏度环境下发射时O型环有失灵的迹象。

确实——75华氏度环境下出现了问题。70华氏度环境下也出现了问题。63华氏度环境下也同样出现了问题。事实上，有证据表明O型环在7次互不相关的发射任务中发生了热损坏。如果观察一下这几次发射时的气温，你会发现很难找出一个规律。通过观察这些数据，你会轻易地相信温度并不影响O型环的运作状态。[注]《委员会报告》指出：“通过比较，O型环‘失灵’这一情况在连接处温度处于53华氏度到75华氏度区间内并没有呈不规则分布。”

问题在于这种比较方法仅研究了24次航天飞机发射任务中7次的数据。人们通过仅关注O型环出现问题的发射任务截短了数据组——这么说听起来很委婉，但言下之意就是他们并没有对所有数据展开研究。这个数据分析错误将导致严重的后果。

因为工程师们仅仅研究“发射失败的情况——如果你看一下先前7次存在问题的发射任务和17次没有问题的发射任务，研究一下当时的气温，你就会发现其中存在明显的关联，这一关联十分重要，但他们却忽略了这一情况”。安·E.滕布伦塞尔和马克斯·H.贝泽曼在给《魔鬼经济学》[注]写的一篇博客文章中提出了上述观点。[注]

你一旦看了所有的数据——把没有发生事故的发射任务也包含在内——你自己就能看出区别。

气温在65华氏度（18℃）以上，20次发射任务中只有3次出了事故。

气温在65华氏度以下，4次发射任务均出了事故。

这是一个依赖数据——虽然出发点极好——导致灾难性后果的典型例子。[注]

表2－1 O型环发生热损坏的发射任务

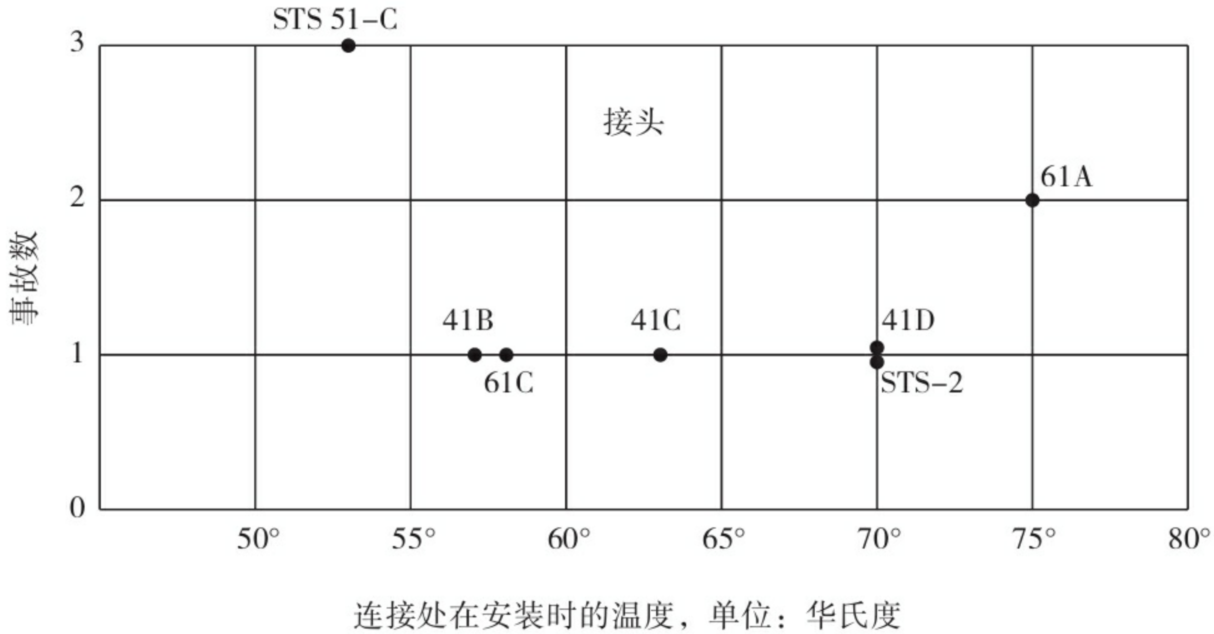


表2－1标出了O型环发生热损坏（即O型环出现腐蚀、漏气、过热等状况）的发射任务。注意该表仅标出了O型环失灵的发射任务，并且以温度记录其分布情况。

表2－2 所有发射任务

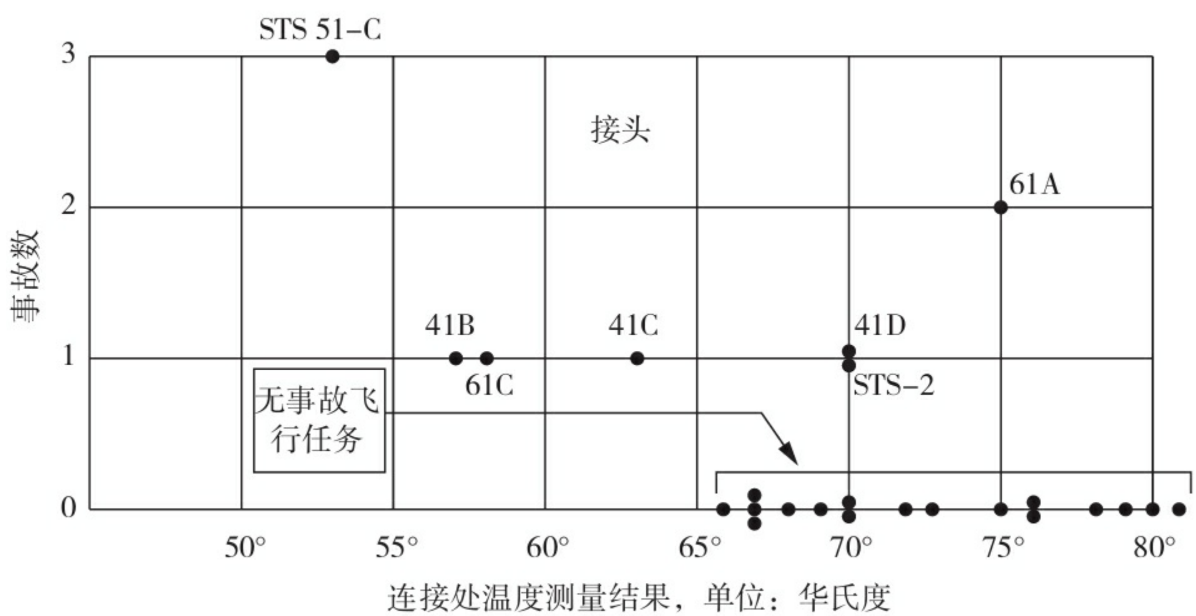


表2－2标出了所有发射任务——O型环失灵和没有失灵都包含在

内。同样，以温度记录其分布情况。

所幸你自己所选择的数据样本可能永远不会陷入最终导致7人殒命、重建费高达17亿美元的航天飞机毁于一旦这样的处境。[注]但不论你在家读报纸还是在公司写报告，每天也都会碰到与之类似的数据问题。我们会在接下来的几页中解释样本是什么，并向你展示如何避免得出错误的结论。

---

1. You should also note that, while Figure 2-1 displays the number of incidents,it does not accurately convey the amount of thermal distress. According to the Rogers Commission Report, the worst “blow-by” (a symptom of distress) occurred at 53 degrees. Looking at this chart, you can only see how many incidents there were— not how serious each one was.
2. 《魔鬼经济学》简体中文版已由中信出版社出版。——编者注
3. Ann E. Tenbrunsel and Max H. Bazerman, “Launching Into Unethical Behavior: Lessons from the Challenger Disaster,” Freakonomics blog, June 1, 2011,accessed April 25, 2015, http://freakonomics.com/2011/06/01/ launching- into- unethical- behavior- lessons- from- the- challenger- disaster/.
4. To be fair, looking at the sample incorrectly was just one of the issues cited by the Rogers Commission Report, which also cited “a faulty design unacceptably sensitive to a number of factors” (chapter IV). The putty used to help seal gaps was also the subject of much debate, given that NASA changed suppliers (the original contractor used asbestos, and later stopped making the putty). Leak checks— meant to ensure the integrity of the putty and O-ring seal— also had the unfortunate effect of creating “dangerous gaps in the putty,” according to Power to Explore. Many articles and books have explored the various reasons for the tragedy, and it is certainly not our intention to state (or even imply) that the sampling error alone caused the disaster. We only use it to illustrate our point. Communication (or lack thereof) was also a key component. In one particularly damning section of chapter V of the Rogers Commission Report, the commission noted that the decision makers “were unaware of the recent history of problems concerning the O-rings and the joint and were unaware of the initial written recommendation of the contractor advising against the launch at temperatures below 53 degrees Fahrenheit and the continuing opposition of the engineers at Thiokol after the management reversed its position . . . If the decisionmakers had known all of the facts, it is highly unlikely that they would have decided to launch 51-L ［the name for the flight］ on January 28, 1986.
5. “Frequently Asked Questions About the Space Shuttle and International Space Station,” Kennedy Space Center website, accessed April 25, 2015,

http://www.nasa.gov/centers/kennedy/about/information/shuttle_faq.html#1.

我们为何需要抽样？

假设你有一盒蜡笔，共100支。你想要知道其中有几支是蓝色的。在这种情况下，你需要做100次观察以收集所有数据。观察即一个简单地看一眼一个单位的动作。

针对一盒100支蜡笔，对每一支蜡笔进行观测以研究这一数据总体是行得通的。对数据总体进行观察是有好处的，因为你无须对余下的数据做假设。

但如果你一年来人际关系处得出奇地好，在生日的时候收到了一个盒子，里面装了100万支蜡笔，这意味着什么？意味着现在你的数据总体为100万。

谁又有时间去观察100万支蜡笔呢？

所幸，还有另一个方案可供选择：使用一点数据分析手段，你依然可以估算出其中有多少支蓝色蜡笔。怎么估算？你可以从数据总体中抽出一个样本。样本即总体中的一部分（并非全部）。一旦你得到了样本，你就可以对总体做推断。（但就如你先前所见，在“挑战者号”航天飞机的例子中，如何选择样本会对数据分析产生巨大的影响。）

在数据分析中，使用样本对总体进行估算是一个常用手段。但你必须非常小心，因为哪怕很小的错误都会迅速放大，因为对数据样本进行观测就相当于对许多其他数据进行了观测。如果你想保证自己的推断正确无误，就还需考虑其他许多因素。最终你还要知道自己所选择的样本对数据总体来说有没有统计学上的意义，这一点我们会在本书第五章中讨论。

如果是这样，结果会如何？

研究数据，得出统计结论的过程称为推断。在所有的统计分析中，如果你想要推断出结果，就必须保证自己拥有针对所要解决问题的正确数据。在本章中，我们会探究两类比较重要的抽样问题，鉴于大家对接收数据已有所了解，我们认为这两类问题应当引起重视。㊟

第一，你需思考自己所研究或收集的数据是否能够代表基础数据总体。这一问题与数据收集方式和收集什么样的数据有关。回到我们讨论过的蜡笔的例子：如果给蜡笔装盒的人在盒子一侧放了较多的蓝色蜡笔——而那一侧正是你抽样的地方，结果会如何？在这种情况下，如果你想要估算盒子里有多少支蓝色蜡笔，你心中所想的数字会比实际上多出许多，因为基础数据并不能代表整盒蜡笔。你所研究的样本向蓝色蜡笔偏倚。

第二，你需考虑针对所收集的数据，分析的对象是什么——你是在分析所有数据，还是分析一部分数据？例如，假设你拥有盒中所有蜡笔的数据，你想要知道多少蜡笔是蓝色的。但在分析过程中，你仅将标签上写着“蓝色”的蜡笔确定为蓝色的。在这种情况下，你只研究了数据中的一部分（蜡笔标签名），这可能会导致你对蓝色蜡笔数据总体的估计和一个观察过每一支蜡笔的颜色，并把水绿色、青绿色、丹宁色蜡笔也观察了一遍的人所得出的结论大不相同。

你必须时常问自己，是否能够准确地将研究发现从样本推广到数据总体？这种推广的可能性称作外延有效性，即把从样品中得出的结论加以推广，从而得出对整个数据总体有意义的结论。

本章的例子着重指出了数据抽样的多种方式——这些方式中，有些

正确，有些错误。为了符合这一章节的主旨，我们在此假定所有和数据抽样相关的错误，都是不经意、都不是故意产生的。这样的分类十分重要，因为有时候人们会出于特定的目的对数据进行抽样，以得出他们想要的结果，即有选择性地对数据进行抽样——我们会在第七章详细讨论这个概念。

正如你所看到的，在分析小数据的时候，抽样是基础。如果抽样出了问题，在解读数据的时候不可避免地会出错。有的数据专家穷尽整个职业生涯，就为了让抽样能够准确并具有代表性，从而为后续的数据分析奠定坚实的基础。这些数据专家身价十分高，因为离开了他们，一切都将崩塌瓦解。

---

1. There are many issues with sampling— these are only two of them. We explore a few more concerns— including sampling error— in chapter 5. That said, this book is meant to be educational and entertaining, and shouldn’t be taken as a comprehensive treatise on every issue that can arise with sampling (or any other topic).

“怪诞”的科学

如果你研究人类行为，那就必须抽样。因为我们的星球上有70亿人，绝无可能将其全部作为样本来进行观察。

所以，如何来选择样本呢？

最为理想化的情况是你有一个足够庞大而且多样化的人群，这些人能够非常准确地代表整个数据总体。不幸的是，从心理学角度讲，这种状态是达不到的，甚至连接近这种状态都难。

某期刊上有文章指出：“人类行为学家总是循环往复地在世界顶级期刊上发表有关人类心理、行为的概括性论断，而这些论断全都是从西方的、教育程度高的、工业化的、富有的、民主的（前五个词英文首字母缩写为WEIRD，意为怪诞）社会中抽样。”[注]

但是，等等，事情貌似更糟了。因为很多研究不仅仅是从那些“怪诞”的社会中抽样得出的，其中有些研究是从最“怪诞”的社会——美国抽样得出的。

有研究人员指出：“美国心理学研究有一个惊人的特点，即研究结论并非基于涵盖面广、差别性大的人群，而是仅仅基于人类总体的一小部分——这一小部分人大多生活在美国。”[注]

美国人口占世界人口不到5%，但在某些研究中却占了样本的68%。“世界上其余的95%人口都被忽略了。”[注]

尽管美国人习惯站在聚光灯下，但你肯定能够明白为何在研究人类的时候把其中大部分忽视掉是很有问题的。尤其因为美国人很难代表世

界其他地区的人群，至少在很多方面无法代表。生活在美国（以及其他WEIRD国度）可能会影响我们感受形象以及处理和金钱的关系等方方面面。[注]事实上，有一名研究者指出，WEIRD国家的人民可能是“在研究智人领域，最不能用来普遍化的子人群”。[注]

稍等，问题还不止这些。

在美国，一项研究表明绝大多数心理学研究“在很长时间以来一直以大学生为研究对象——尤其是以上心理学课程的本科生为研究对象。这样的情况已经持续了近50年”。[注]

在某期刊中，有2/3的美国研究样本为心理学本科学生。[注]

在某种程度上，使用这样的样本是有道理的。心理学本科生在每所大学中都无所事事，他们都愿意挣些零花钱。因此在很多情况下，这样的学生是研究人员的“上佳之选”。[注]

除了上述特点之外，心理学本科生根本不适合作为研究样本。因为他们的平均年龄低于美国国民的平均年龄。下面举一个例子说明这会带来什么不同。随机抽取一个美国本科生，“他们同西方国家以外的人相比，做研究样本的概率高出4000倍”，[注]这可不仅仅可以用“怪诞”来形容了。这会导致你研究的很多心理学现象得出错误结论。

---

1. Joseph Henrich, Steven J. Heine, and Ara Norenzayan, “The Weirdest People in the World?,” Behavioral and Brain Sciences 33 (2010): 61.
2. Jeffrey J. Arnett, “The Neglected 95%: Why American Psychology Needs to Become Less American,” American Psychologist 63, no. 7 (October 2008): 602.
3. Arnett, “Neglected 95%,” 602.
4. Henrich et al., “Weirdest People,” 6.
5. Henrich et al., “Weirdest People,” 19.
6. John M. Grohol, “Psychology Secrets: Most Psychology Studies Are College Student

Biased," PsychCentral blog, last reviewed August 26, 2010, accessed April 25, 2015, http://psychcentral.com/blog/archives/2010/08/26/ psychology- secrets- most- psychology- studies- are- college- student- biased/.

7. Arnett, "Neglected 95%," 604.

8. Grohol, "Psychology Secrets."

9. Henrich et. al., "Weirdest People," 3.

抽样不一定越大就越好

如果抽样数目很大但不正确，这样的抽样不一定就能保证得出较好的结果。你可以研究世界上任何一个心理学本科生的行为，但这并不能表示你就能够以此推而广之得出全人类的行为模式。

拿美国国家体重控制注册中心（NWCR）做例子，该中心在宣传单上自称在美国境内开展了一项针对长期成功瘦身并维持的调查，对超过1万人的瘦身数据展开追踪。但样本数最“大”并不表示该样本就是好的，就好像去拉斯韦加斯最大的自助餐厅用餐并不一定让你吃得健康一样。的确，你餐盘多放点不同的食物的确能让你吃到健康食品的概率上升，但你必须继续把数据挖得更深一点。如果你挖掘一下美国国家体重控制注册中心发表的有关减肥的研究，你会发现其中很多报告的数据中，女性比男性人数多很多。[注]这并不值得惊讶，因为在那个中心注册的人员中，80%为女性，只有20%为男性。[注]

是否这个情况就能让他们的研究无效？并非如此。

是否这就表明他们的抽样无效？并非如此。

但这的确能够限制将其所得出结论推广至更大范围人群的减肥效果上。尽管人数很多，但这些个体依然属于特定的群体——参与了减肥，收到了减肥效果，并且自愿参与这项研究，而且比较个性化。这仅仅是一个例子，说明为什么你不能根据数据想当然地做出结论，无论抽样数目有多大都是一样。

---

1. “Research Findings,” NWCR, accessed March 31, 2015, http://nwcr.ws/Research/published%20research.htm.
2. “NWCR Facts,” NWCR, accessed March 31, 2015, http://nwcr.ws/Research/default.htm.

系上安全带

《洛杉矶时报》的一则头条新闻可能会让所有子女尚幼的家长停下手边的事，再多看一眼。

“儿童高脚凳致伤率在7年内上升了22%；如何才能保证自家孩子的安全。”㊟

乍看之下这个数据十分可信，毕竟《洛杉矶时报》是全国最大、声誉最好的报纸之一。这篇文章是基于一篇发表在经过同行评议的期刊《临床儿科》的研究之上。文章的第一作者是国家儿童医院研究学院伤害研究政策中心主任。该医院在美国国立卫生研究院资助的医院中排名前十。㊟

此项研究的数据来源于美国“全国电子伤害监督系统”（NEISS），该系统从医院收集病人数据。上面的样本集没有从儿科医生那里收集数据，没有从家长那里收集数据，而是仅仅从医院收集数据——而且是从一部分医院收集数据。㊟换句话说，得出的结论仅仅反映了一部分和儿童高脚凳有关联的伤害情况。

在这章的前半段，也就是“挑战者号”航天飞机的部分，我们看到了专家们是如何仅仅研究所掌握的一部分数据的情况的。在这个案例中，专家似乎研究了NEISS数据库中的所有数据。但问题在于尽管研究了所掌握的所有数据（去医院就诊的受伤情况），他们依然没有掌握所有数据（所有的受伤情况）。在这种情况下，研究人员没有掌握所有的数据，很难回答人们关心的核心问题。

公正地讲，如果你研究儿童高脚凳致伤率，NEISS的数据似乎是一个合理来源，而且不论是文章的作者还是《洛杉矶时报》的记者，都很

清楚数据的来源。但这并不能改变我们作为成熟的数据接收者取得信息的方式。

例如，假设在研究过程中，因为高脚凳致伤的儿童数量总体在减少，但受伤后去医院就医的儿童人数在上升，结果会怎么样？在这种情况下，受伤儿童的总数在减少，但家长决定带儿童去医院的比例在增加。为什么呢？也许由于出现了一则广为流传的新闻，越来越多家长担心孩子脑震荡。也许出现了一项新的研究，越来越多儿童医生建议家长带孩子去医院。也许的确受伤人数在下降，但受伤程度比以前更为严重。如果你不提出这些问题，你将永远不知道上面新闻的说法有多少可信度，也不会知道如何正确地解读它。

再说得清楚一点，我们不是说不想让孩子更安全。我们都会保护孩子（甚至过分保护）——我们会在买厢式旅行车的时候事先研究各类汽车的碰撞试验排名。(注)我们在这里只是说你应该知道数据的来源。

如果你对孩子负责的话，记得在孩子坐高脚凳的时候为他们系上安全带。因为绝大多数受伤事件是在孩子试图站立或从高脚凳上爬下来的时候发生的。(注)

---

1. Deborah Netburn, “High Chair Injuries up 22% in 7 Years; How to Keep Your Baby Safe,” Los Angeles Times, December 9, 2013.
2. “There Is Something Remarkable Happening Here,” Nationwide Children’s website, accessed April 25, 2015, http://www.nationwidechildrens.org/ hospital- overview.
3. The data was only from some hospitals. According to its website (“National Electronic Injury Surveillance System (NEISS),” United States Consumer Product Safety Commission website, accessed April 25, 2015, http://www.cpsc.gov/en/ Research-- Statistics/ NEISS- Injury- Data/), NEISS offers “a national probability sample of hospitals in the U.S. and its territories” for injuries related to consumer products. The total number of injuries is then estimated, based on this sample of approximately 100 hospitals (the sample size changed throughout the years). Is that a large enough sample size? Is it a representative sample?While we have no reason to believe that the sampling was misguided (and we’d need another chapter to fully explore it), we do want to point out that this is yet another instance in which sampling has the potential to

significantly impact the everydata in our lives.

4. Check out SaferCar.gov. Just keep in mind that some rankings are relative to all other vehicles— while some results “can only be compared to other vehicles in the same class and whose weight is plus or minus 250 pounds of the vehicle being rated,” http://www.safercar.gov/Vehicle+Shoppers/5-Star+FAQ#one. In other words, a three- star- rated SUV could theoretically be safer than a five star compact car.

5. Rachel M. Kurinsky, Lynne M. Rochette, and Gary A. Smith, “Pediatric Injuries Associated with High Chairs and Chairs in the United States, 2003– 2010,”Clinical Pediatrics 53, no. 4 (2014).

我们是第1名，也是第58名！

根据《普林斯顿评论》上派对学校（Party School）排名，雪城大学位列第1名。[注]

根据《美国新闻与世界报道》上国立大学排名，雪城大学位列第58名。[注]

尽管两个排名都针对同一所学校，但是从完全不同的数据中抽样，提出不同的问题。这完全是个仁者见仁、智者见智的问题。《美国新闻与世界报道》看的是毕业率、院系实力以及其他的标准；派对大学排名是基于有关饮酒、毒品、大学生联谊会/女学生联谊会生活以及与此类似因素的学生调查问卷结果得出的。

当然，我们能够在这一章节中不厌其烦地讲为什么我们不能比较这些排名（方法不同等等）。你肯定会问不同的排名方法是如何得出结论的。

其实很简单：你可以基于抽样的数据和提出的问题，对同一个人、同一个地点、同一个事物做完全不同的研究。

---

1. “Syracuse Named Top Party School,” Syracuse University website, accessed July 20, 2015, http://www.syracuse.com/news/index.ssf/2014/08/syracuse_named_top_party_school_princeton_
2. “Syracuse University,” U.S. News and World Report website, accessed March 16, 2015, http://colleges.usnews.rankingsandreviews.com/ best- colleges/ syracuse- university- 2882.

不接受自拍

上周你吃了多少次垃圾食品？

上个月你看了多久电视？

你开车时到底开多快？

当你询问人们有关他们自身的问题时，会面对得到错误信息的风险。人们不总是诚实的。我们有各种偏见。我们的记忆也远称不上完美。凭借着自陈式数据，你会假定十之八九的事适用于所有人（事实上不是）。你会觉得人们能够客观理解自己的行为（事实并非如此）。（想要对自陈式数据的风险有一个更为深入的了解，请在英国科学基金会Brainwaves的博客上阅读《自陈的危险》一文。）[注]

潜在的不确定性所导致的结果便是自陈式数据可能并不可靠。《安全研究杂志》上的一项研究显示，大多数人认为自己开车水平比普通人高。[注]另一项研究报告显示，人们会把自己的身高报得偏高，把自己的体重报得偏低。[注]当你让人们观察自己的时候，这些情况都可能发生。

有时候，自陈式数据是唯一可以获得的数据。有时候，自陈式数据仅仅是更加容易获得。并不能说自陈式数据就是不好的，因为这部分取决于数据的背景，部分取决于数据的收集方式，部分取决于问题的询问方式。如果你想做一个成熟的数据接收者，自陈式数据只是你需要留心的一个方面而已。

---

1. Rob Hoskin, “The Dangers of Self- Report,” British Science Association Brainwaves website, March 3, 2012, accessed September 1, 2015, http://www.science brainwaves.com/ the-dangers-of- self- report/.

2. Allan F. Williams, “Views of U.S. Drivers About Driving Safety,” Journal of Safety Research 34, no. 5 (2003): 491– 494, doi:10.1016/j.jsr.2003.05.002.

3. S. Connor Gorber, M. Tremblay, D. Moher, and B. Gorber, “A Comparison of Direct vs. Self- Report Measures for Assessing Height, Weight and Body Mass Index: A Systematic Review,” Obesity Reviews 8, no. 4 (July 2007): 307–326,http://www.ncbi.nlm.nih.gov/pubmed/17578381.

选举总统与人口普查

美国的下一任总统可能会把胜利归功于抽样。

你知道，总统必须获得选举团的大多数选票才能获胜。这些选票一部分基于美国人口普查，注而美国人口普查依赖抽样来获得准确的数字。

因为根据美国宪法，每10年要对美国境内所有人口做一次普查。注

观察从美国东海岸到西海岸的全部人口几乎是不可能的。人口普查局在官网上写道：“人口普查过程中，有一些人口没有被计算进去。”注《时代》杂志报道：“1990年人口普查遗漏了大约800万人——其中大多数为移民和城市少数族群，最终人口普查局重复统计了400万美国白人。”注

为什么这很重要？除选举外，联邦政府按照人口普查数据分配资金、支持社区。一个地区人越多，得到的支持就越多。

根据女参议员洛蕾塔·桑切斯的说法，阿纳海姆市在1990年人口普查的时候，少计算了7000多人，损失了150万美元联邦资金。这笔钱能够让阿纳海姆面貌大变——或者说能够让任何城市面貌大变。“这笔钱能够让我们的街道更安全，我们能够为无家可归人员建造房屋，我们能够给失业者做培训。”注

跟这一章节其他例子不同，这里讲的不是错误抽样，不是错误解读数据，而是说明抽样的影响（这影响常常为人所忽略）。有人能够因为数据抽样坐上总统的宝座——街道会由于抽样而导致巡逻警察数量减少，我们认为这是你必须知道的事。

1. Each state gets a certain number of votes in the Electoral College, based in part on the number of seats it has in the U.S. House of Representatives. These seats are allocated based on the results of the U.S. Census. For example, after the 2010 census, 8 states gained members and 10 states had fewer members.
2. “Decennial Census of Population and Housing,” United States Census Bureau website, accessed March 9, 2015, https://www.census.gov/ programs- surveys/ decennial-census/about.html.
3. “What Is the 1990 Undercount?,” United States Census Bureau website,accessed March 9, 2015, https://www.census.gov/dmd/www/techdoc1.html.
4. Amy Sullivan, “Why the 2010 Census Stirs Up Partisan Politics,” Time magazine, February 15, 2009. On a related note, to make up for the people who aren’t counted (or are counted more than once), the U.S. Census Bureau conducts a post-enumeration survey. This survey samples households, then compares that data with the original census data. The Census Bureau uses this sample data to develop adjusted population counts. But not everybody wants an adjustment—especially if it’s going to increase the number of people who might vote against you in the next election. As the Time magazine article noted, “In very general terms, Republicans would prefer to err on the side of undercounting and Democrats would prefer to err on the side of overcounting.”
5. Congressional Record, H1602, March 24, 1999, http://www.gpo.gov/fdsys/pkg/CREC-1999-03-24/pdf/ CREC- 1999-03-24- pt1- PgH1602.pdf.

取其精华，去其糟粕

想一下最近全美掀起的去麸质狂潮。新闻头条都在讲最近不吃含麸质的食品的饮食狂潮。麸质是一种给予面团韧劲的蛋白质。但在美国，到底有多少人真的不吃含麸质食品？

根据调研公司NPD团队的一次市场调研结果，有超过29%的美国人尝试不吃含麸质食品。[注]加上最近热火朝天的去麸质运动影响，去麸质食品的潜在市场估计会超过4400万人。[注]（聪明的读者可能发现全美29%的人口要远远大于4400万，这种不同很可能是由于不同的收集数据方法导致的——还不算那些试着不吃含麸质食品和真的去买不含麸质食品人群的不同。）

现在我们来把上面的数据和患有乳糜泻的人数做比较。所谓乳糜泻是一种和小肠无法吸收麦麸相关的潜在疾病。根据美国国家乳糜泻防治基金会的数据，每133个人中有1人患有此疾病——全美的人数大约为2400万，不到总人口的1%。[注]

这个例子中，对全国人口展开研究和对人口中与此事相关的子集（患有乳糜泻或对麦麸过敏的人群）所受的潜在影响展开研究，其结果大不相同。《赫芬顿邮报》的戴维·卡兹博士解释道，有些人不吃含麸质食品会感觉舒服一点，还有一小部分人吃不吃含麸质食品“可能是生死攸关的大事”。“对于除这些人之外的人来说，不吃含麸质食品充其量只是跟风而已。”[注]

许多人罹患乳糜泻这一事实是一个严重的问题。但我们的观点是基于抽样的人数以及提问的标准，你可以对同一问题（“多少人不吃含麸质食品？”）得到两个截然不同的答案。如果你问全国所有人中有多少

人试着不吃含麸质食品，答案是29%。如果你咨询乳糜泻防治团体全国多少病人需要吃不含麸质的食品，你得到的答案是不足1%。与此类似，通过研究全国人口得出的不含麸质食品对健康影响不大这一结果如果应用到全国不足1%会因食用含麸质食物而丧命的人身上，将会是十分危险的。

结果真是天壤之别。

---

1. “Is Gluten- free Eating a Trend Worth Noting?,” NPD Group website, accessed April 25, 2015, https://www.npd.com/perspectives/ food- for- thought/ gluten- free- 2012.html.
2. Kathie Rowell, “Growing Number of People Choosing Gluten- free Lifestyle,”Shreveport Times, April 13, 2015.
3. “Celiac Disease: Fast Facts,” National Foundation for Celiac Awareness website, accessed April 25, 2015, http://www.celiaccentral.org/ celiac- disease/ facts- and- figures/
4. David Katz, “Is Gluten- Free Just a Fad?,” Huffington Pos website, last updated September 24, 2011, accessed April 25, 2015, http://www.huffingtonpost.com david- katz-md/ gluten- free- diet_b_907027.html.

填空

有时候，该有数据的地方没有数据。

我们来看一下全美第三大学区芝加哥公立学校的例子。这个学区发布了大量关于其学生优异表现的数据，[注]也用数据测定学生的成长情况，作为教师、校长考评的一部分，也作为学校可靠性的一个因素。[注]

但如果你研究一下2014年美国西北测评协会（NWEA）学业进展测评（MAP）的数据，你没准会发现有一些政府特许学校并没有提供信息。[注]事实上，如果你往下拉看完8322行Excel表格文件的话，你会发现有不少没有提供信息的例子——单元格是空白的。

为什么？有一些情况下，政府特许学校要么成立时间较短，要么班级规模非常小（有些学校符合这两个因素），因此它们没有提供数据便讲得通了。

但也有其他情况，就是有些政府特许学校已经开了5年甚至10年之久，可以推测数据肯定是有的，所以应该只是这些学校没有上报而已。芝加哥一所学校的校长在《芝加哥太阳时报》的一篇短文中提到："有人跟我说那些政府特许学校并非一定要'参与'MAP评估……（芝加哥公立学校）允许一些政府特许学校不参与这项评估，以此维持普通公立学校的可信度。"[注]（公立学校似乎没有不参与这项评估的特权。）

很有可能缺失的数据并没有对总体比较结果产生巨大的影响。[注]但当你读到"芝加哥公立学校比政府特许学校进步更快，在阅读教学方面尤其如此"这样的头条新闻时，[注]你便能够看到基于数据编造的故事是怎么改变我们的观念、影响我们的行为的。

如果潜在数据没有经过准确的抽样，那就仿佛建一座房子，却少了几块混凝土地基。可能这并不要紧。但如果缺失的地基位于错误的位置——或缺失太多地基，那整座房子将会倒塌。

---

1. “School Data,” Chicago Public Schools website, accessed April 25, 2015, http://cps.edu/SchoolData/Pages/SchoolData.aspx.
2. “NWEA Measures of Academic Progress (MAP),” Chicago Public Schools website, accessed April 25, 2015, http://cps.edu/SchoolData/Documents/NWEA_MAPFactSheet.pdf.
3. We downloaded the spreadsheet on March 8, 2015, at http://cps.edu/School Data/Pages/SchoolData.aspx.
4. Troy A. LaRaviere, “Drop CPS’ Reform Strategy: CPS Neighborhood School Growth Outpaces Charters,” Chicago Sun Times, September 1, 2014.
5. And we’re certainly not going to get on a soapbox to talk about the pros and cons of charters versus public schools or the wisdom of standardized testing. Although if you want to learn more, a guest post on the Brain Pickings blog is an interesting read: Anya Kamenetz, “5 Reasons Standardized Testing Won’t Slow Down,”Penelope Trunk blog, last reviewed March 5, 2015, http://education.penelope trunk.com/2015/03/05/5- reasons- standardized- testing- wont- slow- down/.
6. Art Golab, Becky Schlikerman, and Lauren FitzPatrick, “CPS Outpaces Charter Schools in Improvements, Especially in Reading,” Chicago Sun Times,October 1, 2014.

缺了什么？

有时候，你无法保证数据抽样绝对正确，因为这样的数据根本不存在。

奥巴马总统在决定发起突袭抓捕奥萨马·本·拉登的时候说：“下这个决断非常困难，部分原因是我们并没有掌握确凿证据。我们只有间接证据证明他会到那里去。”[注]

换句话说，他没有掌握全部数据。

这种事在政治、商业、学校、家里都常常发生。你拥有一个数据组，但你知道这个数据组并不完整。（这样常常依然比你自己认为完整的数据组更好，虽然它并不完整。）

因此，下一次你想要做决定的时候，问一下你自己：你希望拥有什么数据？什么样的数据可以让你改变主意？我们并不是说你要在下结论之前拥有所有数据，因为这样的话你就永远不会去赴初次约会或者尝试吃寿司。但你对先前未知的东西知道得越多，你就能过得越好。

---

1. Barack Obama, “The Raid That Killed Bin Laden,” 60 Minutes, May 9, 2011. http://www.cbsnews.com/news/ president- obama-on- the- raid- that- killed- bin- laden/.

做一个成熟的数据抽样接收者

1.明白基于错误样本得出结论将会导致的后果。样本是否能够代表总体？例如，那些接受问卷调查的人是谁？所选的数据是否基于你想要研究的关键结果，研究这个“样本”是如何影响分析结果的？

2.问自己：哪些数据能够最为恰当地回答所提出的问题？比如，“挑战者号”调查小组一度只研究O型环出问题的发射任务，因为研究结果显示这些事故在一定气温范围内时有发生，所以也许这个数据并不是能够解决问题的上佳之选。如果调查组把注意力集中在所有发射任务上，他们可能会发现O型环问题在温度较低的情况下更为频发。如果要回答一个有关人类行为的问题，你想了解哪些类型的人？或者，换个说法，受访者的答案是如何影响特定分析结果的？

3.在报纸上读到一则新发现或新研究报道时，问你自己：这个结论是研究了哪些数据得出的？受访者是谁，或者这项研究是基于何人开展的？**当你看到“在一个非科学的调查中……”或“领先的”________，等诸如此类的话，你可以将其视为危险信号。**你可以问自己5岁的孩子下一任总统是谁，然后将这个“非科学的调查”结果发表。“领先的”这类词听起来不错，但很难量化（你可能会问“领先于什么？”）。

4.留心自陈式数据——问人们有关他们自己的职业，看什么电视节目，或行为如何等问题与观察并记录人们行为相比会简单很多，但自陈式数据并非总是最准确的。当你看到那些由研究对象提供的数据时，记住上面的话。

5.最后要记住，在很多统计工作中，对数据的一部分抽样或对数据的总体抽样并无好坏之分。抽样是一个强有力的工具，能让我们在研究

总体不可行（或并不推荐这么去做）的时候了解到情况。你不要被误导，认为一定要研究所有数据才行。事实上，给数据抽取一个样本会非常有用。在有些情况下，研究数据的子集所得出结论的确会非常有意义而且非常合适。（AC尼尔森公司创立者阿瑟·查尔斯·尼尔森说过：“如果你不相信抽样，那下次抽血的时候，让护士把你全身的血全抽光算了。”[注]）在其他的一些情况下，我们所研究的结论是由我们分析特定数据组所得出的。所以研究数据的一部分还是全部并不重要，解读结果的方式以及结果的意义才是关键所在。

---

1. As quoted in Ratings Analysis: Theory and Practice by James Webster, Patricia Phalen, and Lawrence Lichty, (Lawrence Erlbaum Associates, 1991), 113.

第三章 红色州为什么变蓝了：平均数及总数——近观概括性统计

你应该知道得克萨斯州的埃尔帕索吧，在那里会举行得克萨斯西部最大的文身、音乐节“得克萨斯对决节”。也许你之所以知道这个地方，是因为史蒂夫·米勒乐队在那里惹上了大麻烦。[注]

不过，如果你从政，你可能会知道埃尔帕索虽然是计选票的地区之一，但那个地方的选票并不总能统计出来。

嗯？

我们来回忆一下2012年总统选举，巴拉克·奥巴马在埃尔帕索县（埃尔帕索的所在地）获得了70%的选票。他的主要对手米特·罗姆尼获得了28%的选票。[注]事实上，奥巴马在得克萨斯州获得了超过300万张选票。

但由于数据的计算方式，导致这些选票并没有给选举结果带来任何不同。

我们都知道没事别惹得克萨斯州的人。所以，到底发生了什么？

答案就是统计。

我们在本书前面章节提到过，美国使用选举团这种选举方式进行选举，而这种选举方式在大多数州施行的是获胜者获得全部选票的方式。在多数情况下，总统候选人只要在某一州获得多数大众选票，就能够获得该州所有选举团的选票。最终，便出现了“红色州”——在那里共和党

候选人获得选举团选票，和“蓝色州”——在那里民主党候选人获得选举团选票。

的确，2012年奥巴马在得克萨斯州获得了超过300万张选票，但罗姆尼获得了超过400万张选票。因此罗姆尼获得了得克萨斯州选举团的全部选票。

选举团是数据合计的一个例子——数据合计是概括性统计的一种，这种统计方法会抹杀数据的多样性，因而常常具有误导性。你也许见到过媒体报道分析红色州和蓝色州可能存在的所有不同点，着重指出了从创造工作岗位到环境保护法到奥巴马医保方案等一系列不同。[注]但我们思考、行动、投票的时候，真的有那么多不同点吗？或者如果我们更深入地研究一下数据，是否会得出不一样的结论？

我们来近距离看一下投票数据，就从密歇根大学的马克·纽曼所总结的2012年投票结果地图（图3–1）来看一下吧。[注]

你把红色州和蓝色州对比着看，会发现有很多不同。例如佛罗里达州是美国东南部唯一一个蓝色的州。但之所以会这样，是因为你看到的只是州一级的合计数据。换句话说，你看到的是选举团竞选方式所得出的获胜即取得全部选票的结果。[注]

尽管地图上的州红蓝分明，但其中依然可能会有很多人投了相反的选票。

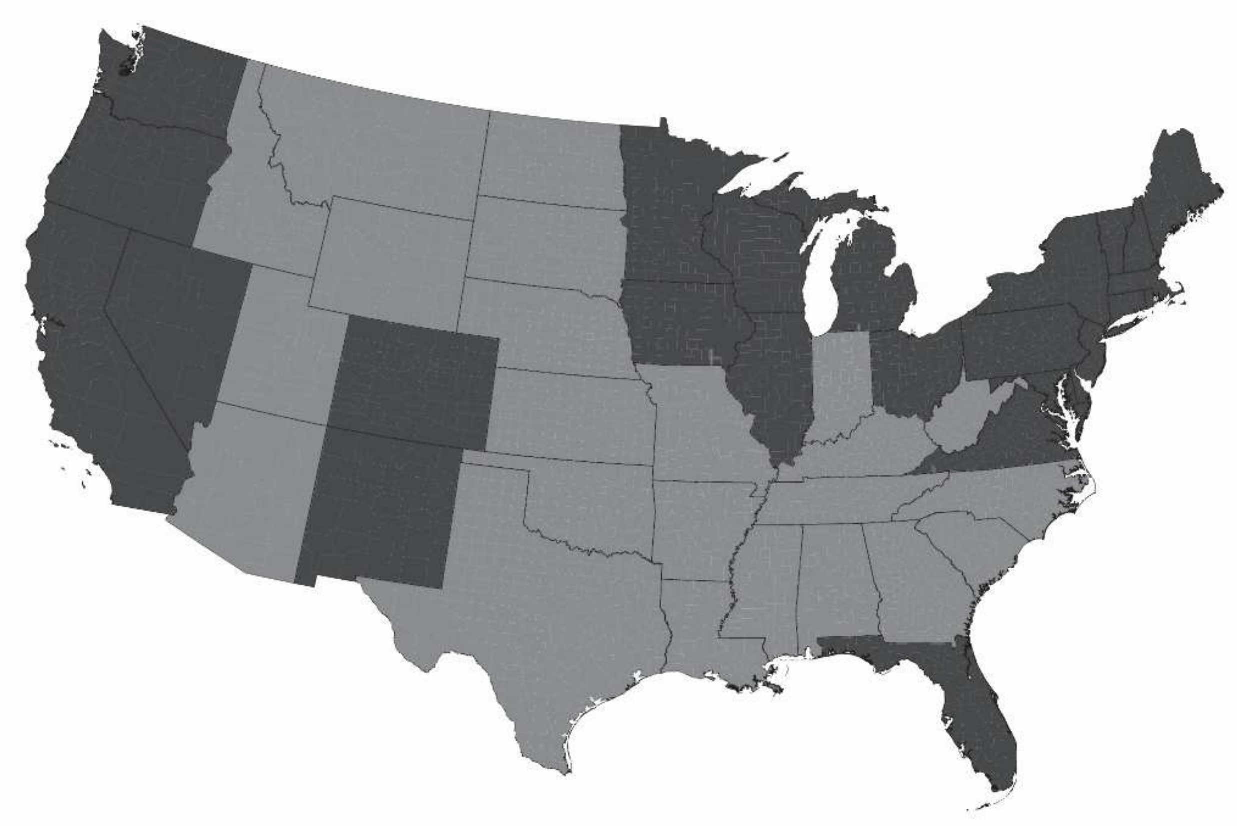


图3－1 州一级的合计数据（浅灰色的是红色州；深灰色的是蓝色州）

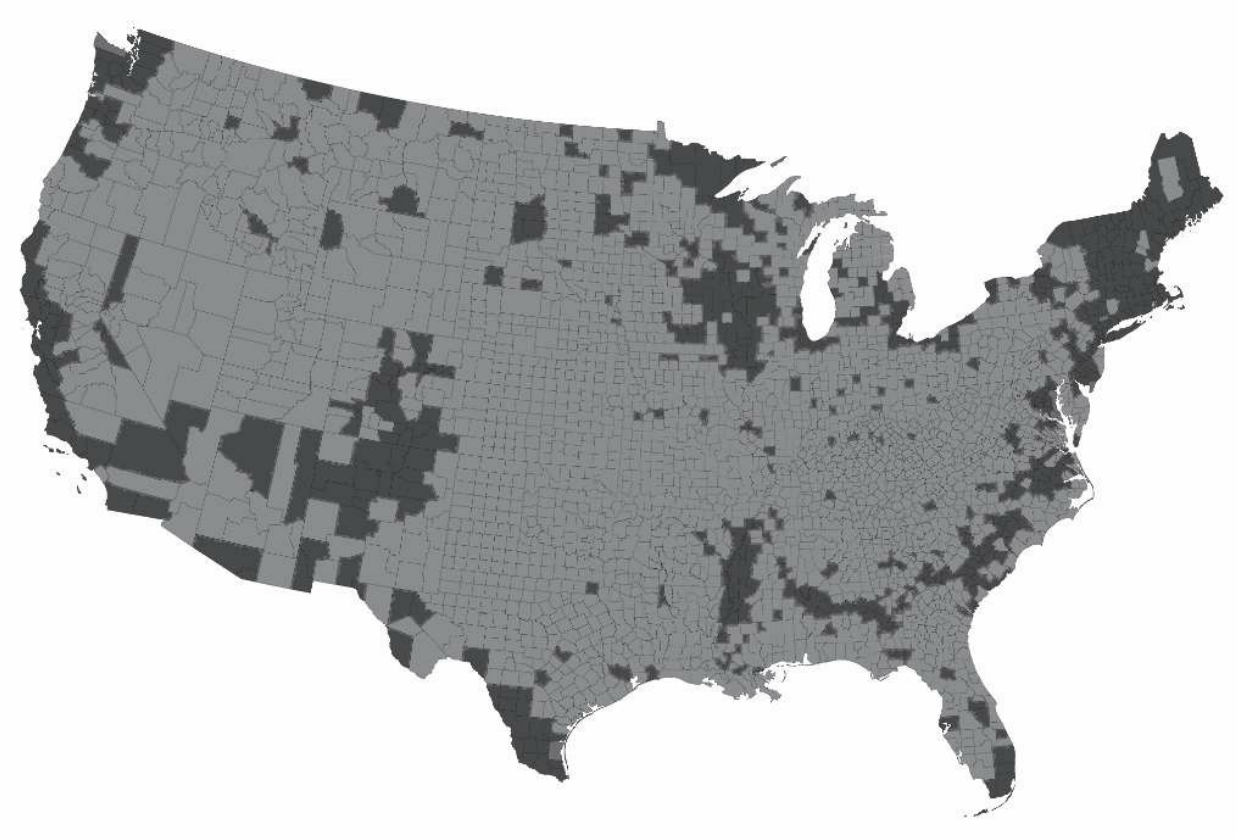


图3－2 县一级的合计数据（浅灰色的是红色州；深灰色的是蓝色州）

纽曼绘制了一系列有关2012年选举各州结果的地图，非常好地体现了这一点。在这张图上（图3–2），纽曼以县为单位描绘了选举结果。靠近一点，你就能看到埃尔帕索县是得克萨斯州西北角上的一个蓝点（深灰色）。

现在，我们看到了以较小单位（县）而非较大单位（州）统计的投票结果地图，该地图看起来就没那么泾渭分明了。除了一些特例（佛蒙特州），大多数州是红蓝相间的。

（我们必须注意，并非所有人都投了民主党或共和党。比如在2012年，有67326名美国人——大约相当于加利福尼亚州雷东多比奇的人口，投了罗斯安尼·巴尔的票。你没听错，就是那个罗斯安尼·巴尔。）

当然，尽管这些以县为单位的投票结果依然只显示了每个县的投票总数（要么民主党，要么共和党）。因此纽曼更进一步，基于投票率，用粉色阴影（我们在该书的地图中将其转换成了灰色）表明每个县为候选人投票的意愿是否强烈。这依然是县一级的数据总计，但我们加上了另一个分解层面——即每个县为候选人投票的意愿是否强烈。在图3–3，纽曼进一步将红蓝界限模糊化，展现了一个更为微妙的政治分布图。埃尔帕索县为深灰色（深蓝）笼罩，反映出奥巴马获得了70%的投票（事实上为69.84%）。

本章节中，三张地图都用了相同的选举结果数据。都总结了投票数，但使用了不同层次的总和。结果，根据数据合计的方法和数量，三张地图显示了投票者的偏好，这样的偏好或多或少有点微妙。

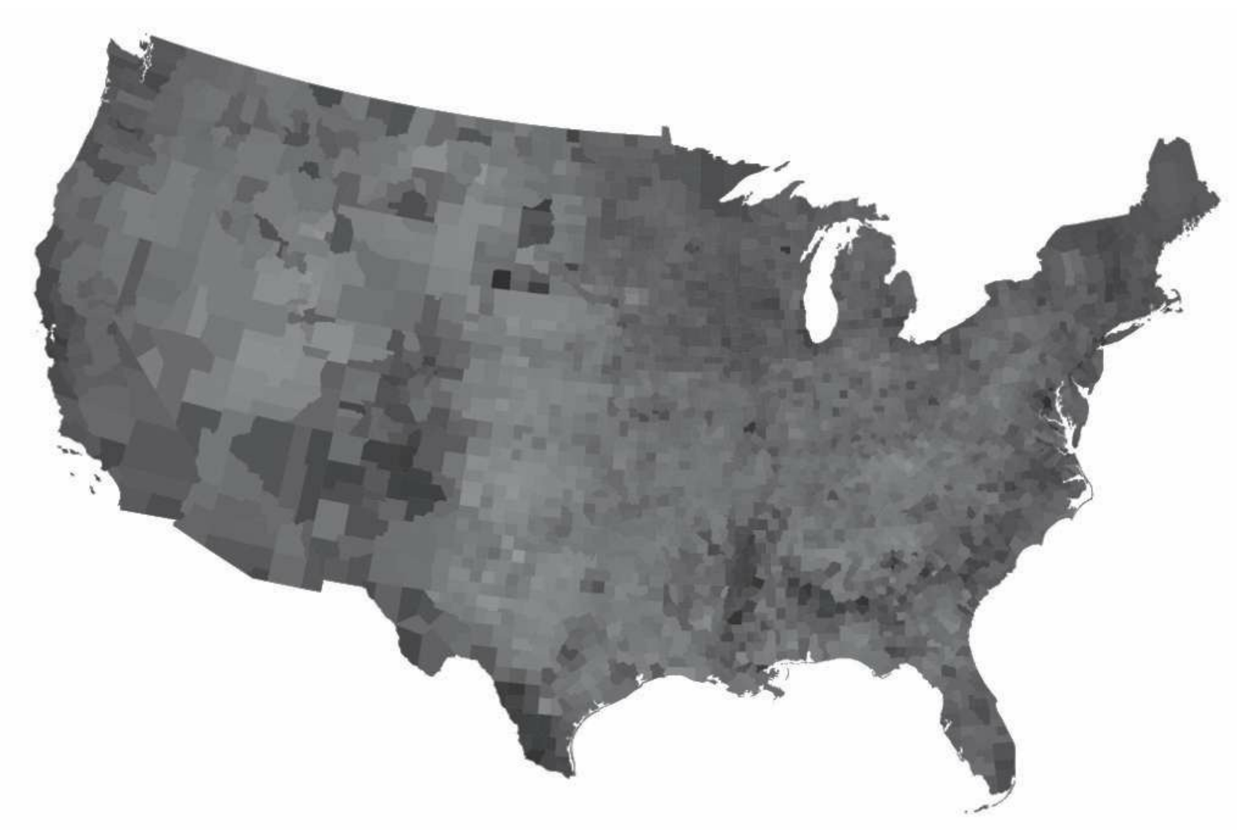


图3－3 加上为候选人投票意愿是否强烈后的县一级总计

数据合计的方式会掩饰重要的不同点。以州为单位的合计数据使得

得克萨斯州成为一个红色州，而以县为单位的合计数据使得得克萨斯州的一些县呈蓝色。在本章中，我们将探索人们总结数据常用的一些统计方法，并展示不同的数据合计方法是如何掩饰基本数据潜在的重要的不同点的。

---

1. “Take the Money and Run,” Steve Miller Band, from the album Fly Like an Eagle (Capitol, 1976).
2. Election Summary Report, November 14, 2012, http://assets01.aws.connect.clarityelections.com/Assets/Connect/RootPublish/elpaso-tx.connect.clarity elections.com/ElectionResults/2012_ElectionDocs/110612/ELECTION_RESULTS_ED.pdf, accessed through El Paso County Elections website(https://www.epcountyvotes.com/).
3. See, for example, “Are Red or Blue States Better Job Creators?,” CNBC, accessed August 16, 2015, http://www.cnbc.com/2015/08/13/ are- red-or- blue- statesbetter-job- creators.html; “Climate Battle Will Likely Divide Red States and Blue States Down a Green Line,” National Journal, accessed August 15, 2016, http://www.nationaljournal.com/ next- america/newsdesk/ climate- epa- regulationobama- states- 20150803; Steve Benen, “ ‘Obamacare’ Thrives in Nation’s Largest Blue State,” the Maddow Blog, accessed August 16, 2015, http://www.msnbc.com/ rachel- maddow- show/ obamacare- thrives- nations- largest- blue- state.
4. All maps are from http:// www- personal.umich.edu/~mejn/election/2012/.Mark Newman, “Maps of the 2012 US Presidential Election Results,” from the personal page associated with the University of Michigan website, updated November 8, 2012, http:// www-personal.umich.edu/~mejn/election/2012/.Text and images used under Creative Commons license: http://creativecommons.org/licenses/by/2.0/.
5. You also see a lot of red, which is surprising given that Obama won, until you realize that the illustration is a map based on the size of the state rather than its population. On Newman’s website (“Maps of 2012”), he offers cartograms that provide a different look at the results. And in case you were wondering, the two states that do not follow a strict winner- take- all approach are Maine and Nebraska, which can split their votes among multiple candidates.
6. Newman, “Maps of 2012.”
7. Federal Election Commission, Federal Elections 2012: Election Results for the U.S. President, the U.S. Senate and the U.S. House of Representatives (Washington D.C., July 2013), http://www.fec.gov/pubrec/fe2012/federalelections2012.pdf.

当心数据缝隙

《全球幸福指数报告》自诩为“全球幸福里程碑式的调查”。[注]该报告为联合国分支机构所发布，每年会有超过百万人次的阅读量，全球主要媒体都会对其进行报道。

报告上，排名第一的是瑞士，再往下看，就好像冬季奥运会颁奖典礼一样列出各个国家的名次：挪威、加拿大、瑞典等国都名列前茅。

把报告文件下拉，你会看到英国位列全世界最幸福国家第21位。这个排名对英国来说还算体面，因为其排在了德国、法国、西班牙前面，但仍然落后于之前曾是其殖民地的新西兰和澳大利亚。[注]

如果你把这份172页的报告翻一遍，就能看出这份报告的研究手段及结果。你能够（也必须）提出一个问题，即研究人员是如何对人口进行抽样并对结果进行解读的。

但吸引我们注意力的是拿国与国进行比较这一概念。一旦将国家看作一个整体，《全球幸福指数报告》中的国家排名取的就是平均值。

记者、研究人员，以及其他人员使用取平均值的方法掌握数据的某些方面，就好像用照相机拍照，能够拍出多个视角，但每张照片只能从一个角度进行拍摄。就好像对数据取总数一样，取平均值是概括性统计中的一种，能够告诉你数据的一些内容，但毕竟只是一种衡量方式，而且这种衡量方式常常带有欺骗性。把所有数据收集起来，接着将其整合为一个值，取平均值（或其他概括性统计方法）暗示了所有基本数据都是相同的，哪怕事实上这些数据并不相同。

英国的“国家统计办公室”（ONS）也收集幸福数据，因此我们可以

对数据进行更为细化的研究，而不是仅仅从《全球幸福指数报告》上看国家的平均值。就像你预期的那样，当你更近距离观察了幸福指数，你会发现英国不同地区的幸福指数不尽相同。拿1到10来打分，不同地区有着不同的分数，如：(注)

北拉纳克郡——7.0分；

英格兰东北部地区——7.2分；

圭内斯郡（威尔士）、东柴郡（英格兰）——7.6分；

埃利安锡尔、奥克尼郡、设得兰群岛（苏格兰）——7.9分。

（当然，就算是上面的分数也只是当地人口的平均分。很可能这些分数会根据特定的周围环境或基于不同性别、收入、年龄而不同。如果你们中有人感兴趣的话，根据ONS的数据，英国的幸福指数平均分为7.3分。）

我们并没有想要将ONS的数据和《全球幸福指数报告》的数据进行比较——我们也不能这么做，因为这两个报告使用的是不同的研究方法，研究覆盖不同的时间段，除此之外还有其他的差异。我们列出这两组数据，只是想要揭示平均值是如何掩盖数据差异的。

英国的每个国民都是不同的。甲壳虫乐队不是滚石乐队。但取一个国家的平均值会无视这些差异，仅从地理角度把所有人都划到一起。

---

1. John F. Helliwell, Richard Layard, and Jeffrey Sachs (eds.), World Happiness Report 2015 (New York: Sustainable Development Solutions Network, 2015),http://worldhappiness.report/.
2. The data comes from the Gallup World Poll, which asks people to evaluate their lives on a scale of 0 to 10 using a tool known as the Cantril Self- Anchoring Striving Scale. The scale is named for Hadley Cantril, a researcher who studied, among other things, people’s reaction to Orson Welles’s famous War of the Worlds radio broadcast. The Cantril scale asks people to think of their life satisfaction as a ladder, and rank their life as the best (10) or worst (0) or

somewhere in between. Hadley Cantril, The Pattern of Human Concerns (New Brunswick, NJ: Rutgers University Press, 1966).

3. Personal Well- Being Across the UK, 2012/13. Contains public sector information licensed under the Open Government Licence v3.0. Permitted under license.“Open Government Licence for Public Sector Information,” the National Archives website, accessed July 9, 2015, http://www.nationalarchives.gov.uk/doc/ open- government- licence/version/3/.

平均数、中位数和众数哪个更可信？

从统计学角度讲，谈到平均值，你必须了解三个概念：平均数、中位数、众数。

所谓“平均数”，即大多数人所谓的“平均值”。想要取平均数，就把所有的数值相加，再除以数据个数即可。

所谓“中位数”，即中间值。如果你将数据以从小到大的顺序排列，中间那个数据就是中位数。（如果数据总数为偶数，那中位数就是最中间两个数据的平均数。）

所谓“众数”，即数据中出现次数最多的那个数。如果没有数据出现次数大于一次，那就没有众数。在另一方面，如果有两个（或两个以上）数据出现次数相同，那将会有多个众数。

我们来看一些数据样本，更直观地了解一下上面的概念。[注]

想象一下，2013年你住在西雅图的切尔西公寓。这座为西雅图世界博览会而造的公寓楼有10个公寓单元。[注]

现在，我们假定其中9个公寓单元为你和你的伙伴所居住（一人一个），而且巧合的是你们的年收入完全相同——5万美元。

一天，比尔·盖茨决定大量裁员，并搬进了第10套公寓居住。他那年的收入为115亿美元。[注]要计算公寓楼中住户年收入的平均数，你先要把每个人的年收入相加：50000+50000+…+50000+…+11500000000=11500450000（美元）。接着除以人数（10），得出平均年收入1150045000（美元）。

要取10个人年收入的中位数，你要把数据从小到大排列，并挑出中间那个数值。

1.50000

2.50000

3.50000

4.50000

5.50000

6.50000

7.50000

8.50000

9.50000

10.11500000000

中位数为50000。（因为我们数据的总数为偶数，所以我们取中间两个数值的平均数。）

现在，我们来计算10个人年收入的众数，你只要计算一下哪个数值出现次数最多即可。

年收入	出现次数
50000	9
11500000000	1

这就很简单了，年收入的众数为50000，因为这个数值在10次中出现了9次。

因此，年收入的平均数超过10亿美元，但中位数和众数都是5万美元。如果你想要计算这座楼住户的年收入，哪个值最为准确？（提示：可以取上面三种方法得出来的数值的平均数！）在这种情况下，取中位数和众数会比较好。（不过，如果你想要估算所得税所产生的影响，那平均数可能更为准确。）

当你观察数据的时候，如何才能知道人们所说的“平均值”是什么？如果你假定“平均值”即平均数，很有可能你是对的，但如果想得到确定的答案，唯一的办法就是询问。

---

1. Thanks to Richard Strausz, a board member of the Detroit Area Council of Teachers of Mathematics— and Mike’s wife’s uncle— for the inspiration for this section.
2. “Chelsea Apartments,” Yateswood.com, accessed August 5, 2015, http://www.yateswood.com/sites/default/files property- files/Chelsea%20Offering_0.pdf.
3. Julia La Roche, “Here’s How Much 10 of the Richest People in the World Made per Minute in 2013,” Business Insider, December 19, 2013, http://www.business insider.com/ what- warren- buffett- makes- per- hour- 2013-12#ixzz3jyUhYnBx.

迈阿密人出生时平均是西班牙裔，死时是犹太人？

按照平均值来看，迈阿密人生下来是西班牙裔，死的时候是犹太人。

这是个玩笑话，但能看出依赖平均值的危险之处。（迈阿密戴德县总出生人口中，60%为西班牙裔，而该县的犹太人只占5%。）[注]

平均值的优点在于将数据中的所有数值都计算进去，并将其简化为一个数值。然而，这个优点也暗含着巨大的危险。如果数据中每个值都完全相同（可以将其想象成一排一模一样的砖块）那平均值可能的确能够准确反映出每一个数值的一些特性。但如果数值在很多关键方面并不相同——很多数据组都不相同——那平均值很可能让高于或低于平均值的数值或同平均值不同的数据组变得含糊不清，还会使一些数据显得与众不同。

试想有两个男人，一个体重150磅（68千克），另一个体重250磅（113千克）。他们的平均体重为200磅（91千克）。但是，这个平均值是否能提供给我们他们个人体重的有用信息呢？在这种情况下，平均值无法准确地回答我们所关心的问题。根据统计学家雷蒙德·布列特的统计，2015年波士顿马拉松赛的平均完成时间为3小时46分钟。[注]但没有任何女性组（布列特将比赛人员进行了分组）参赛者跑得比平均成绩要好。如果我们比较关心女性马拉松运动员的成绩，那上面的平均成绩又能告诉我们什么呢？

平均值误导我们的另一个方式便是仅仅抓住了数据的一个侧面。比如，《嘉人》杂志（澳大利亚版）上的一篇文章列举了体重相同的6位女性——154磅（70千克）。[注]但当你看她们的外形时，会发现她们体

形各不相同，身高也从不到5英尺到6英尺多（145~186厘米）不等。平均体重没有涵盖任何其他数据。

最后，就好像我们在本书中涉及的所有数据概念，平均值并不能孤立地看待。平均值只有在有其他数据支撑的时候才有用。如果你的样本无法反映出数据总体的情况，如果你有目的地选取数据，或者如果你的数据有其他问题，那你所得出的平均值将具有误导性。

---

1. “March of Dimes Peristats,” March of Dimes website, accessed July 10, 2015,http://www.marchofdimes.org/peristats/ViewTopic.aspx?dv=mt®=12086&top=2&lev=0&slev=6; “2014 Greater Miami Jewish Federation Population Study: A Portrait of the Miami Jewish Community,” Greater Miami Jewish Federation website, accessed July 10, 1015, http://jewishmiami.org/population study/ and Census Bureau data.
2. Raymond Britt, “Boston Marathon 2015 Statistical Analysis,” Competitor.com, updated April 22, 2015, http://running.competitor.com/2015/04/photos/ boston- marathon- 2015- statistical- analysis_127026.
3. Julie Adams, “Who Weighs the Aussie Average?,” Marie Claire.com.au website,accessed July 10, 2015, http://nicolepartridge.com/wp-content/uploads/2012/05/MC05_FEAT_70kg.pdf.

为什么副市长会比市长赚得多？

这儿有个例子可以说明基本数据是如何影响平均值的。

在美国，市长/镇长的平均工资为62000美元。

副市长/副镇长的平均工资为83000美元。[注]

副市长/副镇长怎么就比市长/镇长赚得多呢？嗯，哪怕只有一个红绿灯路口的小镇，都会有镇长，那样的小镇长拿的工资不高。《南怀德比纪事报》上的一份研究发现大部分小镇长的收入都低于1万美元。[注]当你计算全国平均值的时候，这些低薪水的官员会把整个数字拉低。

另一方面，拿薪水的副市长/副镇长是只有大城市才能够负担得起的职位。城市越大，薪水越高。比如纽约市有四个副市长，每个副市长的工资能到20万美元。[注]

你研究市长/镇长工资的时候，你看的是所有城镇的数据。如果你研究的是副市长/副镇长的工资，那你所研究的范围就会缩小到那些自治市（而且很可能拿更多薪水）。因此，如果你只看那些有副市长/副镇长的城市，你可能会发现市长/镇长的平均工资比副市长/副镇长高。但如果你把设有以及没有设有副市长/副镇长职位的城镇都计算进去，你最终得出的数据——第一眼看上去——会完全出乎意料。

这是另一个统计概念相互重叠、相互影响的例子。如果你看的是平均值，顾名思义，你研究的是一个特定的数据组。如果你比较平均值，而且这些平均值来自不同的样本集，那样本集的不同会在平均值中显现出来。记住，平均值并不比基本数据更胜一筹。[注]

来谈谈工作与平均值，你知道如果待在一个单位不动，赚的钱会比较少吗？根据《福布斯》的说法：“在同一公司任职超过两年，一辈子下来平均会少挣大于等于50%的钱。”[注]

这是为什么？根据《福布斯》杂志的说法，因为如果你待在同一家公司，考虑到通货膨胀等因素，你的净收入平均来讲只增加1%，但跳槽的人平均收入会增加10%~20%，甚至更多。把这一点考虑进去，如果你换个几份工作，那你工资单上的收入将会明显增加。

我们不是说你应该跳槽。平均值仅仅说明了做这项研究的小组的工作成果，但具体的情况会因人而异。不过有时候平均值能为你所期望达到的目标指点一二。

如果你想要从事播音行业，你可能会想看一下电视新闻主持人的工资待遇。根据广播电视数字新闻协会与霍夫斯特拉大学的年度调查显示，2015年，主持人的“平均”工资为83800美元。这个收入并不差，但如果你看一下工资的中位数，你会发现仅仅只有65000美元。[注]得出这个结论会有多个解释（包括最高工资高得吓人，为87.5万美元，这个工资能够大幅度提高平均工资，但不会对工资中位数产生多少影响，甚至不产生任何影响）。但设想一下如果你想要去学校读新闻专业，当你听到“新闻主持人平均工资为83800美元”和听到“新闻主持人工资的中位数为65000美元”这两种说法中，哪种说法能让你更倾向于选择走新闻行业这条路？

---

1. “Simply Salary: Mayor,” Simply Hired website, accessed July 10, 2015, http://www.simplyhired.com/salaries-k- mayor-jobs.html; “Simply Salary: Deputy Mayor,”Simply Hired website, accessed July 10, 2015, http://www.simplyhired.com/salaries-k-deputy- mayor-jobs.html. And yes, we have some concerns about the accuracy of this data for numerous reasons, including that it’s only from online job listings, and that the source doesn’t defineaverage (we’re assuming it’s mean,but can’t be sure). That said, we’ve chosen to include this example because it illustrates a much- needed point about how sample sets can impact averages.

2. “How Much Mayors Make, Langley Council Plans Review of Mayor’s Salary,”South Whidbey Record, May 22, 2011, http://www.southwhidbeyrecord.com/news/122414919.html.

3. “Mayor, Office of the (OTM),” NYC Citywide Administrative Services website,accessed July 3, 2015, http:// a856- gbol.nyc.gov/gbolwebsite/390.html. In fact, the first deputy mayor (who is second in command) has a higher salary than the mayor— although this could be because the mayor receives other benefits,including the use of Gracie Mansion on the city’s Upper East Side.

4. As one of John’s colleagues noted, this could be seen primarily as a selection issue. His point was that a more accurately selected sample would not be misleading, because we could either look at just a sample that had mayors and deputy mayors, or we could look at all cities and put in a zero as the deputy mayor’s salary for cities that don’t have a deputy mayor. Either way, you would likely get an answer showing that the deputy, on average, makes less than the mayor.

5. Cameron Keng, “Employees Who Stay in Companies Longer Than Two Years Get Paid 50 Percent Less,” Forbes website, June 22, 2014, http://www.forbes.com/sites/cameronkeng/2014/06/22/ employees- that- stay-in- companies-longer- than-2- years- get- paid-50-less/.

6. Bob Papper, Research: RTDNA Salary Survey, Radio Television Digital News Association website, July 13, 2015, http://www.rtdna.org/article/research_rtdna_salary_survey#.VaUV00XFTbU.

如何评估学生的成绩？

平均值如果从最纯粹的角度来看，每个数据都得到了平等对待，但这并不总是你想要得到的结果。

试想你是个英语教师，你想要给学生最终成绩。一年来，学生一共经历了三次小测试、一次期中考、一次期末考。

如果你想公平对待每一组数据，你可以直接把所有考试成绩相加，并除以5（考试次数），得出学生的最终成绩。但可能你会觉得期末考试应该比期中考试占的比重更大些，而期中考试又应该比小测试占的比重更大些。

这种情况称为加权平均。加权平均值顾名思义：你基于重要程度给每一个值加以权重。比如你基于对每项考试对学生成绩重要程度的判断，认为三次小测试应该占总分的10%，期中考试占总分的20%，期末考试占总分的50%。现在，假定你在深思熟虑后，确定了每一项考试所占的百分比，那最终得出的加权平均值将更为准确地体现学生的表现。

侧重于使用加权平均值可能与简单（未加权）平均值有所不同，因为从定义上讲，加权平均值对某些数据更为重视。当你想到平均值，试着确定这个平均值到底是简单平均值还是加权平均值。如果是加权平均值，弄清楚这个平均值是如何加权的，并找出那些数据更为重要。在我们刚才讨论的成绩例子中，很明显期末考试更为重要，对平均值的影响也较大，老师（可能）有意对这些测试加以不同权重。在其他情况下，加权可能是无意为之，或看起来不那么明显。

平均值的平均值

平均值将数据整合起来，但如果基本数据已经经过了整合，会发生什么事？

如果你看新闻的话，可能看到过选举之选举——又称整合选举结果，媒体（或其他机构）将不同的选举结果整合为一个容易理解的数字。BBC、Real Clear Politics（美国一政治新闻网站）新闻网以及其他机构会定期提供这样的信息。㊟

现在，你已经了解平均值是基于其组成部分得出的，这些部分越相似，平均值就越能够体现各个值的情况。但在整合选举结果中，你所使用的数据可能在表面上看起来很像，但都由其各自的收集、解读、整合方式所得出。对于一个成熟的数据接收者来说，这样的情况会带来诸多疑问。比如：

所有选举结果是否在整合的时候受到平等对待，或者其中一些是否由于样本大小、时效性、历史准确性等因素被赋予了更大的权重。

每个单独的选举都有误差值。这些误差值如何对整合选票产生影响？（如果的确产生影响的话）取平均值并不会消除基本选举的不确定性。

各个选举的不同点是什么？（调查问卷的问题、样本大小，以及其他选举方法）这些不同是如何影响各个选举结果以及整合数据的？

重点是什么？重点就是尽管整合选票可能（有时候）比单次投票结果更为准确（该结论是普林斯顿选举联盟以及其他机构得出的），但是一个高质量的调查问卷和一大堆质量参差不齐的调查问卷相比，能更好

地告诉我们竞选人的立场。(注)

1. “UK Politics, How Poll Tracker Works,” BBC News website, April 10, 2015,http://www.bbc.com/news/uk- politics- 13248622; Real Clear Politics, http://www.realclearpolitics.com/.
2. Sam Wang, “On the Track Record of Simple Poll Aggregation,” Princeton Election Consortium website, October 24, 2008, http://election.princeton.edu/2008/10/24/on- the- track- record-of- simple- poll- aggregation/.

警惕数据中的异类

除非你处理的是完全相同的数据组，即所有数据的值都一样，否则你的数据中总有一些值和平均值不同。

比如，我们这章讲到的英国幸福指数，我们指出英国的幸福指数为7.3。但英国不同地区的幸福指数从7.0到7.9不等。

这些数据看起来相互吻合。这便是你在调查数据总体的时候期望看到的事。

但有时候，你会发现有一些特定的观察资料与其他材料无法吻合。也许有些数据远远高于（或低于）其他数据，或者也许某个数据和其他任何数据放在一起，都显得格格不入。

这些反常的事物被称为异类。注国家橄榄球联盟球员是异类；14岁就从大学毕业的孩子是异类；那你公司最差，只完成了倒数第二名1/3业绩的销售人员呢？也是个异类。

你在观察平均值的时候，需要留心这些异类。因为你即将发现这些离群值对平均值会产生十分巨大的影响。就好像在清咖啡里面加入奶油。尽管咖啡的含量依然达到95%，但几滴奶油将大大改变咖啡的外观。

难点在于离群值没有一个确定的判断标准。有些经济学家认为理论上离群值就是大大偏离中位数的那个数值，但从实践上讲，这种判断非常主观，而且不同人有不同的解读方式。注这就是为何统计学家花费大量时间逐例分析数据，以判断其是否是离群值。

所以，是什么导致离群值的出现？有时候，仅仅是因为出了差错。可能有人在往空白表格上填数字的时候，把数字顺序搞错了，把4.9填成了9.4。也许试管受到了污染，导致细菌数远远高于正常值。错误难免会发生。

有时候，离群值是出现不正常情况的危险信号。1998年马克·麦奎尔为圣路易斯红雀队打出了70个本垒打，这个情况看起来不同寻常。对于没有使用类固醇药物的人来说，的确不正常。10年后，麦奎尔承认在创造纪录的那个赛季服用了药物，从而证实粉丝和数据学家的怀疑是正确的。

最后，你在阅读或听新闻的时候，要记住有些故事之所以能够成为新闻，只因其报道的是离群值。老生常谈和那些（极为）非比寻常的事相比，常常会显得没有那么激动人心。

---

1. 异类（outlier）在统计学上称为离群值。——译者注

2. For example, some statisticians and economists look for three or four standard deviations (which is a statistical measure of how spread out the data is) as an indicator of an outlier.

去掉最高分和最低分有道理吗？

如果你关注奥林匹克运动会的话，你可能熟悉人们试着应对离群值的方法——直接将其抛弃。例如，在跳水、体操，以及其他运动项目上，运动员的得分是去除最高分和最低分，再将其他裁判的分数相加，再取平均值所得出的。

这种策略——名为截尾取平均数——能够防止因为一个裁判的偏见或个人喜好影响最终结果。根据一篇研究2000年奥林匹克运动会跳水成绩的文章，这种截尾取平均数的方法不止影响了一块奖牌的归属。㊟

然而，截尾取平均数——拿这个方法来处理可能存在的离群值——真的有用吗？你可以问一下自己，如果有超过一名裁判偏向于某个运动员，会出现什么情况？奥林匹克运动会常用的评分方式仅仅去除最高值和最低值。再考虑一下，截尾取平均数这一方式将最高值和最低值当作离群值看待，而不管其是否真的是离群值。这样的评分方式真的公平吗？

现在就出现了问题，即是否最高分和最低分——不论其是否是离群值——是偏见的标志。的确，国别的偏见可能存在——研究人员发现“绝大多数裁判较不是自己国家的运动员而言，会给自己国家的运动员更高的分数”。㊟但是我们来看一下中国的跳水裁判，他在2000年奥运会期间，给中国跳水运动员的平均成绩比非中国跳水运动员要高1.48分。看上去有偏见是吧？但当研究人员分析数据的时候，发现其实该裁判从他给出的分数来看“显然是最没有偏见的”。这怎么可能？因为这个中国裁判给中国以及非中国跳水运动员打出的分数，平均来看比其他裁判都要高。中国跳水运动员水平的确出色，事实上，他们的平均分比非中国跳水运动员要高出1.44分。因此，研究人员观察这位中国裁判打的

分数，发现同其他裁判给本国跳水运动员加的分数相比，事实上这位中国裁判给中国跳水运动员加的分数相对较少。在这种情况下，把这位中国裁判的分数去掉还有道理吗？

---

1. John W. Emerson and Silas Meredith, Nationalistic Judging Bias in the 2000 Olympic Diving Competition, August 22, 2010, http://www.stat.yale.edu/~jay/EmersonMaterials/MathHorizons.pdf. The specific event in which the outcome may have changed was women's 10-meter platform, which the authors explored in: John W. Emerson, Miki Seltzer, and David Lin, "Assessing Judging Bias: An Example from the 2000 Olympic Games," American Statistician 63, no. 2 (2009):124– 131.
2. Emerson and Meredith, Nationalistic Judging.

总统办公室的离群值？

当然，有些离群值并非错误或危险信号——它们是完全有效的数据。以美国历史为例，如果你观察美国总统的在职时间，你会发现大多数总统的在职时间为1460天或2921天（误差为1天），分别对应4年和8年任期。但根据统计学家罗伯特·W.海登博士的分析[注]，有44%的美国总统在职时间大于或小于上面的时间，使其在职时间成为离群值。每次只要总统于在职期间逝世（意味着没有完成其任期），他的在职时间便成为离群值——接替其位置的人的在职时间也将成为离群值。

因此，你如何处理离群值？你将其平等对待，将其和其他数据放在一起，让平均值因之偏斜，还是将其完全忽视？有没有折中的办法？

这要视情况而定。并不存在通用规则，因为情况不同，离群值并不总是能够轻易辨别。比如，有些家长可能会有因为自己小孩的身高排第35个百分位，就认为其身高是离群值。其他家长除非自己小孩排第5个百分位，不然并不对此关心。总之，当你观察平均值，必然会看到有些数值比平均值高，有些数值比平均值低。[注]

关键在于你必须观察数据，判断离群值对你想要回答的问题产生了多大的影响。

这便引出了康伍德公司。

---

1. Robert W. Hayden, “A Dataset That Is 44 Percent Outliers,” Journal of Statistics Education 13, no. 1 (2005), www.amstat.org/publications/jse/v13n1/datasets.hayden.html.
2. An exception being if every value in the data set is identical.

掩盖信息的代价为10亿美元

反垄断法出台以来数额最大的裁决：10亿5000万美元，该案件和离群值息息相关。

康伍德烟草公司以阻碍其发展为名起诉另一家烟草公司——美国烟草公司。[注]康伍德公司的数据专家对美国各个州进行分析，并宣称美国烟草公司对康伍德公司市场占有率产生了不利的影响。[注]

问题出在了这项分析把华盛顿哥伦比亚特区——相比较而言，该地区的市场极小——包含了进去。这就意味着哪怕康伍德公司的产品销售量产生了很小的变化（仅仅在几家店里积压）都会大大改变市场占有率。

分析数据后，可以明显地看到华盛顿哥伦比亚特区和其他48个州相比十分不同（不包括阿拉斯加和夏威夷）。研究反倾销的赫尔伯特·郝文坎普教授称之为“明显的离群值”。[注]但专家并未排除离群数据，而是将其包含在内，使得其他数据因之发生偏离，得出了一个不被其他数据所支持的结论。正如同郝文坎普所说：“原告聘请的专家完全无视了数据中明显的‘离群值’。”[注]

如果排除了离群值——确实可以证明应该排除这样的离群值，那所得出的结论中，康伍德公司的市场占有率将发生明显上升。但是，由于极端的观测方式，导致最终结果显示康伍德公司的市场占有率发生了下降。

如果你排除了一个数据值，导致结果产生了巨大的变化，那这个数据值很可能是离群值。在一个良好的统计模型中，应该能够任意排除一

个数据，统计结果不会因之发生显著变化。这是你寻找离群值时必须考虑到的。

---

1. Conwood Company was purchased by Reynolds American, Inc., and changed its name to American Snuff Company, LLC, effective January 1, 2010.
2. You can read more about the case here: Benjamin Klein and Joshua D. Wright,“Antitrust Analysis of Category Management: Conwood v. United States Tobacco,” November 10, 2006, http://www.justice.gov/sites/default/files/at/legacy/2006/12/01/219951.pdf.
3. Herbert Hovenkamp, The Antitrust Enterprise: Principle and Execution (Cambridge,MA: Harvard University Press, 2008), 81.
4. Hovenkamp, The Antitrust Enterprise, 81.

你是否比一般人更为优秀

一般美国人有以下特点：

每天睡眠时间超过8.7个小时。[注]

体重约为82千克（男性89千克，女性75千克）。[注]

每年喝20.8加仑（79升）啤酒。[注]

每年开车21688千米（希望不是在喝了那么多啤酒之后开的）。[注]

一周淋浴6次，但只洗4次头。[注]

当下的工作已经做了4.6年。[注]

所以，你比一般人更优秀吗？如果有人告诉你，美国有55%的人认为自己比一般人更聪明[注]，而且大多数人觉得自己比一般人颜值更高[注]，而且最近一项研究表明，有93%的人说自己比中等水平司机开车开得更好[注]，你会感到惊奇吗？也许加里森·凯勒在描述乌比冈湖的时候说得没错："在那里女人们都坚强，男人们都帅气，孩子们都优于常人。"[注]

从统计学上讲，93%的司机水平都超过中等水平是可能的。中等水平从定义上讲是你数据中的中位数。但这项研究并没有说93%的美国司机开车水平较高。仅仅说有93%的司机认为自己开车水平较高。

我们在这里看到的很可能就是虚幻优越感的一个例子。所谓虚幻优越感是一种能够解释为何大多数人觉得自己优于其他人，即优于一般人

的认知偏差。(注)

这个问题为什么重要？

如果你觉得自己比普通司机开车开得好，你是否会凭借自己的“车技”飙车或做其他危险的事？

如果你觉得自己赌博水平比一般人水平高，会不会在牌桌前逗留更久（下更大的注）？

如果你觉得自己比一般人更聪明，你会不会应聘超出自己能力范围的工作？（男性朋友们，把头抬起来——和女性相比，男性会更倾向于高估自己的智力水平。）(注)

你能够成为世界上最优秀的统计学家，但如果你没有捕捉到这些认知偏差，那你解读数据的能力会因之受到影响。

---

1. American Time Use Survey, Bureau of Labor Statistics website, last modified September 30, 2014, http://www.bls.gov/tus/charts/sleep.htm.
2. Christopher Ingraham, “The Average American Woman Now Weighs As Much As the Average 1960s Man,” Washington Post Wonkblog, June 12, 2015, http://www.washingtonpost.com/news/wonkblog/wp/2015/06/12/look-at- how- much- weight- weve- gained- since- the- 1960s/.
3. “By the Numbers: What Americans Drink in a Year,” Huffington Pos website,June 27, 2011, http://www.huffingtonpost.com/2011/06/27 americans- soda- beer_n_885340.html.
4. Nicholas Carlson, “If You Drive Fewer Than 9,480 Miles per Year, It’s Cheaper to Take an Uber Everywhere Than to Own a Car,” Business Insider website, September 17,2014, http://www.businessinsider.com/is- uber- cheaper- than- owning-a- car- 2014-9.
5. Grace Gold, “Do Americans Shower More or Less Than the Rest of the World?,”Yahoo! Health website, February 20, 2015, https://www.yahoo.com/health/do- americans- shower- more-or- less- than- the- rest-of-111508762932.html.
6. Employee Tenure in 2014, Bureau of Labor Statistics website, September 18,2014, http://www.bls.gov/news.release/tenure.nr0.htm.

7. Paul Ausick, “Most Americans Are Smarter Than the Average American,”24/7 Wall Street website, May 18, 2014, http://247wallst.com/economy/2014/05/18/ most- americans- are- smarter- than- the- average- american/.

8. Melissa Dahl, “Most of Us Think We’re Hotter Than Average, Survey Says,” NBC News website, September 8, 2010, http://www.nbcnews.com/id/39044399/ns/ health- skin_and_beauty/t/most-us- think- were- hotter- average- survey- says/#.VcO_OflViko.

9. Ola Svenson, “Are We All Less Risky and More Skillful Than Our Fellow Drivers?,” Acta Psychologica 47, no. 2 (February 1981): 143– 148.

10. “Podcast,” A Prairie Home Companion website, accessed September 1, 2015,http://prairiehome.org/listen/podcast/.

11. If you want to learn more about these biases, pick up a copy of Daniel Kahneman’s Thinking, Fast and Slow (New York: Farrar, Straus and Giroux, 2013).

12. Adrian Furnham, Joanna Moutafi, and Thomas Chamorro- Premuzic, “Personality and Intelligence: Gender, the Big Five, Self- Estimated and Psychometric Intelligence,” International Journal of Selection and Assessment 13 (March 4,2005): 11– 24, doi: 10.1111/j. 0965-075X.2005.00296.x.

如何成熟对待数据总和、平均值、离群值

有没有做好准备使用数据总和、平均值、离群值来做更好的决定？这儿有5件事，你现在就能做：

第一，了解什么是**概括统计**，什么不是。很多人觉得因为概括统计能够反映一组数据，那它就能够反映出数据的一切。其实并非如此。概括统计仅仅是一个标准，仅仅从一个维度衡量数据组。就像我们看到的红色州和蓝色州的例子，概括统计会掩盖基本数据的不同点。

第二，**理解所呈现的是哪一种平均——是平均值，是中位数，还是众数。**大多数人听到了“平均”两字，会认为讲的是平均值，但并非总是这样。有三种平均，而且各不相同。其中有一些更容易产生偏离。比如，基于平均值，世界上平均每人有少于两条手臂。（大多数人有两条手臂，但有些人只有一条，有些人没有手臂，因此平均值就被拉下来了，假定只有非常少的人有超过两条手臂。）当有人说到了平均，你要知道自己讨论的到底是哪个概念。

第三，**试问“什么的平均数？”**因为平均值里兼有多个数据值，每个数据值都会影响到最终结果。这便是你可以运用所有在这本书中学习到知识的时候。数据是否体现了样本？你是在看平均数的平均数吗，每一个平均数是否都有自己的特性？除了这些问题，还有很多问题需要问。

第四，**看是否所有数据都被平等对待。**有些平均数只是加权平均数，在这种平均数的计算过程中，有些数据被赋予了更大的权重。比如，有些选举计票使用了加权取平均数，以此来反映到达投票年龄的成年人的真实人数。如果使用正确的话，加权取平均数是一个可行的统计

学工具，但是你必须知道加权是否存在以及加权的方式，从而让自己成为成熟的数据接收者。

第五，**辨别离群值，并理解离群值给平均数带来的影响。**有些离群值是数据组中完全有效的组成部分。其他时候，应当排除极端数值以得出你所问问题的正确答案。常言道，一粒老鼠屎坏了一锅粥。并非每个离群值都是老鼠屎——但你必须注意，离群值会使结论发生偏离。[注]

---

1. Okay, so this isn’t a perfect comparison. A rotten apple literally spoils the bunch by emitting ethylene, a gas that makes the fruit around it ripen more. An outlier doesn’t typically change the data around it— it simply affects the conclusions that you draw from that data. But even though the mechanisms differ, the point is the same— one apple/outlier can change the results.

第四章 使用苹果手机的人更聪明？正确理解关联性和因果性

任何一个注重脸面的家长都会告诉你，想要让小孩子在班上出类拔萃，得对其施加很大压力才行。

这就是为什么我们会让孩子参加智力测试，早早地就让他们参加培训班（在世界范围内，考前辅导、培训、咨询这一块的市场价值超过540亿美元）(注)，并给他们报任何能够想象得到的培训项目。

结果显示，我们完全可以省下这笔钱和时间的。因为我们做过研究——现在，我们来告诉你如何让你家小孩变得更聪明。

最新数据显示，聪明人拥有以下特点：

戴眼镜。（美国在线）(注)

使用苹果手机。（美国有线电视新闻网）(注)

是共和党人。（皮尤研究中心）(注)

听电台司令的歌。（《华尔街周刊》）(注)

睡得晚。（《时尚先生》）(注)

是左撇子。（《纽约客》）(注)

较能喝酒。（《今日心理学》）(注)

以上每一项特征都被引用，认为其与智力有着特殊的联系。因此如果你真的想要在自家厢型旅行车上贴上一张“优等生荣誉之家”的贴纸，显然你需要做的就是给你家孩子配一副眼镜，买一个苹果手机，让他们看几部罗纳德·里根的演讲，放几首电台司令的歌，半夜前别让他们睡觉，让他们用左手做事，并且开始学喝酒（当然，要等他们到能喝酒年龄才行）。

我们是疯了吗？

没有。我们只是阅读了很多经过统计分析，却似乎得出错误结论的研究以及媒体报道。尤其是一些将关联性和因果性搞混的报道和文章，这些文章有时候会潜移默化地将读者误导到错误的结论上。

有必要指出上面的结论有两个问题：第一，有一些原始的科学研究有时候会将关联性和因果性搞混。但你在日常生活中比较可能碰到的是报纸上的文章和其他媒体以错误的方式报道可信的科学研究发现。我们见过许多案例，其中就有新闻将某个研究发现报道为因果性，尽管基础研究报告上标明那个发现和所导致的结果仅仅具有关联性。

从统计学角度，我们能够找到两件事之间千丝万缕的联系，比如戴眼镜和高智商之间的联系。这类联系——数据间存在的某种关系——称为关联性。然而，我们将用这章内容继续讨论仅仅存在这类统计学上两个事情之间的关系，并不能表示两者之间存在真实的、有意义的联系。关联性和因果性不同。这实际上是人们解读数据时犯的最常见错误。但是别担心——在这一章节中，我们会深入研究人们是如何将关联性误认为是因果性的，以及人们为什么会犯这个错误，接着提供给你一个工具，来帮助你认识哪种小数据才是需要相信的。

---

1. GSV EDU, Education Sector Factbook 2012, accessed April 21, 2015, http://gsvadvisors.com/wordpress/wp-content/uploads/2012/04/ GSV- EDU- Factbook- Apr-13-2012.pdf.

2. “People Who Wear Glasses Are Smarter, Study Claims,” AOL website, July 1, 2014, http://www.aol.com/article/2014/07/01/ people- who- wear- glasses- are- smarter- study- claims/20923473/.

3. David Goldman, “Smarter People Use iPhones— Study,” CNN Money website, January 22, 2015, http://money.cnn.com/2015/01/22/technology/mobile/ iphone- smart- study/.

4. “What the Public Knows About the Political Parties,” Pew Research Center website, April 11, 2012, http://www. people- press.org/2012/04/11/ what- the- public- knows- about- the- political- parties/.

5. Marisa Taylor, “Books and Music That Make You Dumb,” Wall Street Journal website, February 27, 2009, http://blogs.wsj.com/digits/2009/02/27/ books- and- music- that- make- you- dumb/.

6. Ned Hepburn, “Smarter People Stay Up Later, Do More Drugs and Have More Sex— It’s Science,” Esquire website, November 21, 2013, http://www.esquire.com/ news- politics/news/a26244/ smart- means- sex- and- drugs- and- staying-up/.

7. Maria Konnikova, “Sinister Minds: Are Left- Handed People Smarter?”New Yorker website, August 22, 2013, http://www.newyorker.com/tech/elements/ sinister- minds- are- left- handed- people- smarter.

8. Satoshi Kanazawa, “Why Intelligent People Drink More Alcohol,” Psychology Today website, posted on October 10, 2010, in the Scientific Fundamentalist,https://www.psychologytoday.com/blog/ the- scientific fundamentalist/201010/why- intelligent- people- drink- more- alcohol.

智能手机＝聪明人？

现在，我们回到智能手机这个问题的分析上。我们来进一步挖掘一下那份研究真正想表达什么，并揭示一些有趣的解释、警告、事实，从而能够对研究成果多一些了解。

我们先从研究苹果手机开始，这份研究调查了美国每一个州的苹果手机使用量和持有学士学位（及其他情况）人口的比例，发现“苹果手机使用率和教育水平呈正相关”。㊟

但这一切仅仅表明苹果手机使用率最高的州正是持有学士学位人口比例最高的州而已，㊟并不意味着你使用苹果手机就更聪明。

关联性不过是一种双变量关系——通俗地说，即两个变量之间存在的关系。双变量关系并不能证明其中一件事导致另一件事发生。试着从这个角度来想：你能够观察到两件事从统计学角度看上去互相关联，但这并不能回答任何你所关心的问题——关系为什么存在，对于数据接收者来说，这种关系又意味着什么。

有许多原因可以解释为什么这种可见的关系并非因果关系。比如，如果你研究一下戴眼镜的人比较聪明这一论点，你会发现原始研究实际上是在描述读书时间长短和近视之间的关联。㊟暂时，我们假定读书时间长的确意味着你比较聪明（这个假设要打上一个大问号，而且和这项研究没有什么关系）。该研究从未说比较聪明的人会戴眼镜，而是指明了在学校上学时间较长和视力好坏的关系。这个区别很大，而且这两者只是互相关联。标题党为了吸引你注意力，把重点放到了视力和智力的关系上，而研究所得出的结论则比较慎重——结论所反映的问题也大不相同。比如，研究中两个因素的因果关系也许恰恰相反——在学校时间

过长可能引起双眼紧张程度上升，从而导致近视，而近视可以用眼镜来矫正。

那睡觉晚的人比较聪明这一观点是否准确呢？嗯，根据原始研究论文，智商最低的一组人平均睡觉时间为晚上11点41分，而智商最高的一组人睡觉时间为晚上12点29分。[注]这篇论文进行了数个统计学分析，发现高智商的人之所以晚睡觉，与其种族、年龄、性别、婚姻状况、父母状况、教育、收入、信仰、工作时间的不同都有关系。在这种情况下，统计分析很明显涉及不止一对双变量关系。尽管如此，许多统计学家依然不会认为其中存在因果关系，而认为其中存在较为复杂的关联性。尽管这项分析研究依然没有清楚地显示晚睡会导致高智商。和之前一样，这种关联性可能是颠倒的。换句话说，一个人脑袋聪明可能导致其晚睡。是不是高智商的人就更容易晚睡，因为他们要做功课，并且看书看到很晚？你可能会看到同样的统计学关系，但这种关系中的因果关系实际上是相反的。

概括一下，如果你想让你的孩子变得聪明，不一定要给他们配眼镜，或给他们买苹果手机。

现在，我们给你节省了一部智能手机的钱，我们继续……

---

1. “Infographic: iPhone Usage Rates by State,” Chitika, Inc. website, January 22,2015, https://chitika.com/insights/2015/iphone-by-state.
2. Assuming the sampling was done correctly, which we would probably question given that they only sampled people during one week of the year— which also happened to be the week of Christmas.
3. “People Who Wear Glasses,” AOL.
4. Satoshi Kanazawa and Kaja Perina, “Why Night Owls Are More Intelligent,”Personality and Individual Differences 47 (2009): 685– 690, https://personal.lse.ac.uk/kanazawa/pdfs/paid2009.pdf.

星巴克旁边的房子升值更快？

人们不仅仅想变聪明。人们想变得更幸福、更健康、更富有……

因此，你会看到这样的标题党——摘自《商业内幕》——“如果你在上下班路上花费超过45分钟时间，那你的婚姻可能会破裂”。[注]

或者这样的新闻标题——摘自网站EliteDaily.com——“裸睡才能有大大的梦想：为何提高生活质量的秘诀如此简单，简单到只要把衣服脱了即可”。[注]

或者这样的新闻标题——摘自《人物》——“住在星巴克旁边能够让你的房子升值”。[注]

我们很愿意住在星巴克旁边，但不是因为星巴克让房子升值，因为这篇文章并没有证实这一点。

文章数据显示星巴克附近的房子5年间升值幅度超过20%，而离得稍远的房子只升值不到17%。[注]我们并不是想对这些数据提出异议，但我们质疑“星巴克让房子升值”这一论点。[注]

可能就因为星巴克把店开在了城镇、乡村的中心——这些地区房价上升的速度较快。

例如，星巴克创意总监兼全球创新发展新型零售部总裁（前不动产总监）亚瑟·鲁宾菲尔德写过一本书，介绍特许经营人如何在每天的工作中学习星巴克选址的经验，他的秘诀中有一条便是在停车场寻找油渍（车流量大的标志）。[注]

也许星巴克坐落的位置旁边会有人行道，而人们喜欢住在有人行道的地方。也许每次星巴克一开，旁边就会开一家苹果商店——这些都是让房地产价格上升的因素。

我们不知道具体原因。而这就是关键所在。

---

1. Kim Bhasin, “If Your Commute Lasts More Than 45 Minutes, You Will Probably Get Divorced,” Business Insider website, June 2, 2011, http://www.business insider.com/ long- commutes- can- cause- divorce2011-6 accessed 4/21/15.
2. Lauren Martin, “Sleep Naked, Dream Bigger: Why the Secret to a Better Life Is As Simple As Taking It Off,” Elite Daily website, August 18, 2014, http://elite daily.com/life/ gotta- free- sleep- free- people- sleep- naked- happier/715351/.
3. Morgan Gibson, “Living Near a Starbucks Will Increase Your Home’s Value,”People’s “Great Ideas” website, January 29, 2015, http://greatideas.people.com/2015/01/29/ starbucks- increases- home- value/?xid=socialflow_twitter_peoplemag.
4. Spencer Rascoff and Stan Humphries, “Confirmed: Starbucks Knows the Next Hot Neighborhood Before Everybody Else Does,” Quartz website, January 28, 2015, http://qz.com/334269/ what- starbucks- has- done-to- american- homevalues/.
5. Rascoff and Humphries, “Confirmed.
6. John Moore, “Starbucks Real Estate Learnings,” Brand Autopsy blog, May 19,2005, http://www.brandautopsy.com/2005/05/starbucks_real_.html; Rubinfeld’s book is Built for Growth: Expanding Your Business Around the Corner or Across the Globe (FT Press, 2005).

还有什么因素可以解释这件事？

上面所有因素——市中心、人行道、苹果商店——有可能是遗漏变量。遗漏变量是造成关联性和因果性不同的主要原因之一。

还记得我们讨论过的双变量关系，即两个变量之间的关系吗？问题在于很多时候变量不止两个。你掌握了两个变量之间的关系（又称依存关系），但实际上却有第三个至关重要的变量被遗漏了。这个变量即遗漏变量。（是的，也可能存在多个遗漏变量。）

星巴克是否让你的房子升值？或者这只是其中一个因素——从理论上来讲，这些因素都会影响到你房子的价格？

有可能星巴克的确让你的房子升值。也可能星巴克只是其他因素（人行道等）的代理。在这个情况下，代理即你认为与其他因素紧密相关但不相同的因素。比如，智商测试仅仅是人内在能力的代理。这样的测试无法衡量你真正的能力，能够衡量的只是你考试的能力，而这样的能力被认为能够衡量智力。有时候代理被（有意无意地）用来补偿一个遗漏变量。有代理总比什么都没有来得好，但很明显代理并不能代替真实数据。

从星巴克的例子中，我们了解到有两组变量之间存在关系。但我们不知道是否存在遗漏变量。在存在遗漏变量的情况下，你不可能知道自己是否已挑拣出所有有意义的关系，掌握了一切能够影响结果的因素。有数不清的实证经济学家穷尽其整个职业生涯都在研究遗漏变量。

只要问一个简单的问题——“还有什么因素可以解释这件事？”，于是你便可以一路寻找遗漏变量了。（有一件有趣的事，和实验数据相比，观测数据中的遗漏变量需要人们多留个心眼。换句话说，如果你观

察两个已经存在的数据——如房价和星巴克门店位置，你可能不会意识到其他影响它们的变量。但如果你是在做实验——比如观察研究生在实验室环境下对奖励的回应——你可以设计实验，从而控制可能出现的遗漏变量。）

最终，你研究关联性的时候，要记住变量可能存在正相关或负相关。你可以将其简单地理解成两个东西朝同一个方向移动，或朝相反方向移动。正相关：星巴克的出现使得房价上升。一个因素上升，另一个因素也随之上升。负相关与此不同：你后院中杂草太多，导致房屋价格下降。在这种情况下，一个因素上升，另一个因素下降。在几乎所有的统计学关系中，我们关心的就是方向。如果我们发现星巴克开在你家的周围，你家的房价下降了，那我们不会得出星巴克与你家房价有关的结论。

我们检验关联性的另一个方法是看符不符合你对于经济学的直觉。比如，如果你看到一项研究显示毒贩人数和房价呈正相关——换句话说，毒贩人数越多，房价越高，我们马上就会怀疑（数据，还有毒贩）。试着分辨不同种类的关联性会有助于你辨别其是否确实是因果关系——以及遗漏变量是什么。

我们有没有让你们觉得无聊？

约翰先前当大学教授的时候，在期末评教时得到了比较公正的分数，随之他对本·施密特教授的性别与评教网产生了兴趣。这个网站上，用户可以输入任意字词，然后看这个字词在“给我的教授打分”的评论中出现多少次。该网站对超过20个学科的男女教授进行评价。（如果你登录benschmidt.org/profGender/#网站，便可以亲身体验。）

从数据角度看，性别差异非常有趣。例如，要是我们在本·施密特教授的网站输入“无聊”一词，会看到一个图表，显示在大多数学科中，男教授和女教授相比，被冠之以“无聊”的人数更多。在工程学领域，女教授被形容为“能力不够”的次数为每百万字18次多一点，而男教授得到类似形容的只有每百万字12次。

不过如果我们想要从一个关键的角度了解不同点（比如老师的性别是否影响他/她被学生形容为“无聊”），我们就要考虑我们观察的关系是否包含了全部，或是否有其他与此相对的因素能够解释我们所观察到的现象。

在这个例子中，想要看出学生是否会因为老师性别而给出不同评价，我们必须确保从一切重要的维度来对相同的教授进行比较。

你应该能预想得到，学生在给予教授主观评价的时候会受一系列因素影响。下面举几个例子：

课堂和科目学习材料有多难？

学生这门课的成绩得了A吗？

这门课布置了多少家庭作业？

这门课的上课时间是不是学生讨厌的（如周五早上8点上课）？

教授是否平易近人？

这门课是一门入门课程，还是一门必修课，或是一门选修课？

现在，虽然有学生在评教的时候最终会受到一系列因素影响这一事实，但这一事实并不足以得出结论。重要的问题是这些“其他因素”是否都会因性别不同而不同。换句话说，是否不同性别的教授收到明显不同的评价，在这些其他因素所描绘的情况下也一一符合？

举一个极端的例子：假定在数学系，女教授总被安排去教难度较大的微积分必修课，而男教授被安排去教非常受欢迎的统计学选修课。如果我们观察到对女教授的评教情况非常不好，那很可能仅仅因为女教授被不均衡地安排去教又难又不受欢迎的课，而学生会更倾向于在结课评教的时候给出负面评价。

这个简单的例子并非想要证明学生给教授评教的不同并非出于教授的性别，但我们肯定可以说，基于这些能够观测出来的简单差别，从统计学的角度并不能得出以上结论。要确定评教和性别之间的真实关系，还需要做更多工作。

为什么重要

如果你想知道数据是如何影响你的生活的，仅仅知道两个变量之间的统计学关系是不够的。我们真正关心的问题是：

如果你看到两个变量之间存在关系，这样的关系是有意义的真实关系，还是只是由于某些原因而存在的虚假关系？

虚假关系的存在有很多原因，不过遗漏变量肯定是导致这种关系常见的原因。寻找因果关系牵涉到我们生活的方方面面。

癌症是由什么引起的？

根据《华盛顿邮报》的一篇文章的说法，非裔和西班牙裔美国学生相对来说不会被界定为“有天赋的”，这又是为什么？[注]

大麻合法化是否安全？

发现所有有效的遗漏变量，寻找关联性和因果性的差别，可以回答数不清的问题，上面的问题只是略举一二。

---

1. Emma Brown, “How Does a Teacher’s Race Affect Which Students Get to Be Identified As ‘Gifted? ” Washington Post website, April 22, 2015, http://www.washingtonpost.com/news/local/wp/2015/04/22/ how- does-a- teachers- race- affect- which- students- get-to-be-identifie-as-gifted/.

穿耐克鞋就能像乔丹一样灌篮？

关联性对于市场人员和媒体来说，是一个强有力的工具，将其与人们渴望在生活的每一个方面都追求更快、更强、更聪明、更性感这种想法相结合，尤其如此。

明星代言体现的都是关联性。试想：

迈克尔·乔丹穿耐克鞋。

迈克尔·乔丹能灌篮。

如果我穿耐克鞋，我能灌篮。

我们自己检验过这样的逻辑，并可以肯定地告诉你，这样的逻辑是不正确的。如果你只有5英尺6英寸（165厘米）高，就算你把迈克尔·乔丹穿过的所有球鞋都买下来，依然可能连篮筐都够不到，灌篮更是天方夜谭。

你不是迈克尔·乔丹。遗漏变量——如他的身高和数不清的练习时间——能够帮助解释为什么他在NBA（美国职业篮球联赛）扣篮大赛中获胜，而你做不到。

某物对某个明星有用，并不意味着它对你也有用。如果认为有用，那就犯了一个典型的把关联性当成因果性的错误。

再举另一个典型错误的例子。杰弗里·布朗是经济学家、伊利诺伊大学商学院主任。在我们采访他的时候，他给我们举了一个例子："假设每个周一早晨，你的狗都会叫。几分钟后，收垃圾的车到了。如果据此就认为狗叫引来了垃圾车，那明显是错误的。在这种情况下，因果关

系可能是相反的——你的狗比你更早听到了垃圾车的声音。很少人会犯与上面的例子一模一样的错误，但他们会在日常生活的其他情况下，在做决定的时候犯类似的错误。”

哪怕你有较多的数据，在很多情况下依然会碰到问题。比如，就像布朗提到的，你可以花上10年观察狗叫和垃圾车的关系，但依然不能证明是狗叫引来了垃圾车。当心这些因果关系反转的情况，确保你不会仅仅因为事件发生的时间来推断其因果关系。

婴儿、洗澡水和波尔多红酒

经济学家艾米丽·欧斯特怀孕的时候，和其他准妈妈一样，她想知道的头两件事便是喝多少咖啡和酒才是安全的。

她不想从朋友和邻居那里寻求一些基于传言的建议，也不想从医生那里获得笼统的建议。她想要的不是关联性。

不过这确实就是许多建议的基础。例如，她在研究咖啡摄入量和高流产率之间的关系的时候，发现了喝咖啡的女性和不喝咖啡的女性之间存在数不清的不同点，“她们自身的不同可能导致了流产率的不同”。(注) 换句话说，她找到了遗漏变量。

幸运的是，作为经济学家，欧斯特知道如何理解数字。她不仅仅听医生的建议或在网上找几篇文章读一下。她发现了解真实情况并非易事。她最终亲力亲为翻阅了几百篇研究报告和原始研究成果，而不是仅仅看别人对研究的解读。的确，她在怀孕期间，偶尔会喝酒，每天也会喝上三四杯咖啡。

哪怕你接受过正确解读数据的训练，而且做起来挺得心应手，但你可能会发现，寻找事实会消耗大量的时间和精力。

很多经济学家致力于观察和实验，而非仅仅依凭理论，花去职业生涯的全部时间评估、思考遗漏变量导致的偏见，社会科学家花上几百个小时分析数据来证明几组变量之间呈现的是关联性还是因果性（还是与之相反）并不稀奇。我们不是在吓唬你——我们只是让你知道挖掘到事物的本质需要经过多少训练，需要花去多少精力。从我们的经验来看，只需知道凭借什么方式，在什么时候对数据提出质疑已经能够让你领先于大多数人了。

1. Emily Oster, “Take Back Your Pregnancy,” Wall Street Journal website,August 9, 2013, http://www.wsj.com/news/articles/SB10001424127887323514 404578652091268307904.

你在搜索引擎页面排第几？

如果你实在没办法辨别出遗漏变量怎么办？

这儿有个例子。如果你开公司，要是你公司网页的浏览量能够翻倍，那你应该会很开心吧。毕竟在谷歌上排名第一的搜索结果浏览量几乎是排名第二搜索结果的两倍。[注]根据行业不同，在谷歌搜索结果排名上只要上升一位，就可能给你带来数以百万计的访问者。

所以如何来提高排名呢？

根据谷歌的说法，他们的搜索结果演算法则基于“超过200种独特的信号（或称为线索），从而猜想用户真正想要搜索什么”。[注]

问题在于谷歌并没有给出这超过200种独特信号的详情——也许谷歌不想因此丧失其竞争优势吧。

你如何处理超过200个遗漏变量？嗯，如果你点开Moz.com网站，你会看到一个图表，上面显示了和搜索引擎排名相关的160个因素。[注]如果你正在寻找提高你在搜索引擎页面排名的方法，那你会发现这很有意思，而且可能非常有用。

但这并不可靠，因为图表上面的因素大多基于关联性而得出。值得称赞的是，在Moz.com网页上，“关联性”一词被使用了12次。[注]在一篇独立的博客文章中，该网站进一步探讨了这个情况，并指出“存在关联性的数据并不（足以）能够给我们指出影响排名的因素”。[注]

有时候，你就是没办法接触到遗漏变量。也许数据不对公众开放，也许被意外毁坏了，或者也许从一开始就没记录在案。在这些情况下，

你可以试着从反向逆推数据，并将呈关联性的数据剔除。只需牢记一句话，尽管你所运用的是能够掌握的最佳数据，但想要证明因果关系依然是一场硬仗。

---

1. “The Value of Google Result Positioning,” Chitika website, June 7, 2013,https://chitika.com/ google- positioning- value.
2. “Algorithms,” Google website, accessed April 20, 2015, http://www.google.com/insidesearch/howsearchworks/algorithms.html. Here, you’ll also find a link to “The Anatomy of a Large- Scale Hypertextual Web Search Engine,” in which Sergey Brin and Larry Page presented Google.
3. “Search Engine Ranking Factors 2015,” Moz website, accessed September 1,2015, https://moz.com/ search- ranking- factors/correlations.
4. “Search Engine Ranking Factors 2015, Expert Survey and Correlation Data,” Moz website, accessed September 1, 2015, https://moz.com/ search- ranking- factors/correlations.
5. Rand Fishkin, “What Do Correlation Metrics Really Tell Us About Search Rankings?,” Rand Fishkin Blog, Moz website, January 14, 2013, http://moz.com/rand/what-do- correlation- metrics- really- tell-us- about- search- rankings/.

烤奶酪的性福生活（我们差点儿将其用作书名）

有一条来自《赫芬顿邮报》的新闻标题写得很好："据调查显示，吃烤奶酪的伴侣做爱次数更多，性格也更好。"[注]（比标题更好的是网址，里面含有"烤奶酪、做爱、翻云覆雨"这类字眼。）

我们知道你在想些什么：这是真的吗？烤奶酪真的那么灵？[注]

根据文章的说法，调查显示，喜欢吃烤奶酪的人中有32%每个月至少做爱6次，而不喜欢吃烤奶酪的人中，只有27%的人做爱能达到这个频率。

哪怕我们无视这项调查的其他纰漏（警惕自陈式数据），也能够很明显地看出，上面新闻讲的两者只是互相关联。人们喜欢吃烤奶酪和性生活存在关系，但并没有证据表明这两件事存在因果关系。因此，这两者是相互关联——而非互为因果。

有一种说法，即单纯从统计学角度说，你无法仅凭观察数据就断定其中存在因果关系。在上面这种情况中，只是没有足够的数据证明是什么导致了云雨时间的延长。会不会还真就是拜烤奶酪三明治所赐？当然可能，但也可能会有上百万个其他的变量。仅仅因为你没有证明两者之间存在因果关系，不能说明两者之间就不是因果关系。

这便是你该如何实践可靠统计。

---

1. Kate Bratskeir, "Grilled Cheese Lovers Have More Sex and Are Better People, According to Survey," Huffington Pos website, April 8, 2015, http://www.huffingtonpost.com/2015/04/08/grilled- cheese- sex- bow- chica- bow- wow_n_7027572.html.

2. The "Cheddar Makes It Better" commercials from America's Dairy Farmers National Dairy

Board were some of our favorites back in the day. You can see one here: https://www.youtube.com/watch?v= ypx8- 7bzLX8.

加利福尼亚的阳光和美属萨摩亚岛上的律师

你知道加利福尼亚的阳光和美属萨摩亚岛上的律师数量相关吗？[注]

你是否还知道全美国保龄球馆的年收入和人均酸奶的消耗量有关？[注]

这些只是我们在虚假关联网站（tylervigen.com）上找到的虚假（即误导性）关联例子中较为好玩的几个。这个网站由哈佛法学院的学生泰勒·维根开设。[注]

虚假关联很有用，因为可以凸显出遗漏变量的存在，并描绘出将关联等同于因果会导致的潜在危险。

我们询问了维根对于不同类型的虚假关联的看法，以及人们如何能够更好地解读它们。“就拿那个老生常谈的冰淇淋销量和谋杀率紧密相关的例子来说吧，”维根说道，“两者的数量都会在夏天上升。其中遗漏变量是炎热的天气，而这样的天气会提高犯罪率是有事实依据的，而且这样的天气也会让人们特别想吃冰淇淋。”在这种情况下，存在着虚假关联，因为存在另一个变量（炎热的天气）——只是人们在展示冰淇淋销量和谋杀率的时候将其遗漏了。

但这儿有另一种虚假关联。我们来看一下加利福尼亚的阳光和美属萨摩亚岛上的律师数量的关联性。“这件事，”维根解释道，“不仅仅是两者间没有明显的联系，而且也没有明显的可以引起这两个变量的第三方变量。”在这类情况下，“这是个纯偶然事件，将这两组数据放在一起其实是巧合。计算机之所以将这两者挑选出来，是因为我给计算机几千组数据处理，这两者之间没有真实的联系”。

根据维根的说法，遗漏变量和毫无联系的虚假变量之间的区别是理解变量的欺骗性中最有趣，同时也是最为难解的部分。“两者都具有欺骗性的取样，但两者的运作模式又截然不同。”

韦尔斯利学院经济学教授考特尼·科尔同意这一说法，并指出“‘虚假关联’十分有趣，但和我们可以想象出关联原因的情况——不论其反映了一个变量对另一个变量产生了一系列轻微影响还是第三个因素可以反映出两个变量之间的关联——相比略逊一筹”。

以上提到的不同点非常重要，在你日常生活中看待互为关联的事物时应牢记于心。自我提问：两者的关联性纯粹只是巧合？还是有遗漏变量在起作用？

作为一个具有批判意识的数据接收者，你可能会想：“是否可能根本不存在巧合？会不会存在一个我们不知道的遗漏变量，从而可以从某种程度上解释加利福尼亚的阳光和美属萨摩亚岛上的律师的联系？”

我们不会说这是不可能的，就好像我们也不会说牙仙是不存在的一样。你的确不能说自己已经掌握了所有的因素，但你通常可以去除那些明显不合情理的因素。（如果你觉得吃烤奶酪三明治可以提高床上生活质量，那你可真比高达奶酪还要性感了。）统计学并非总是完美的，但它提供了一个框架，让人们可以以科学的方式评估数据。

---

1. Tyler Vigen, “Spurious Correlations— Sunlight in California Correlates with Lawyers in American Samoa,” Tyler Vigen website, accessed April 20, 2015,http://tylervigen.com/view_correlation?id=30444.
2. Tyler Vigen, “Spurious Correlations— Total Revenue Generated by Bowling Alleys (US) Correlates with Per Capita Consumption of Sour Cream(US),” Tyler Vigen website, accessed April 20, 2015, http://tylervigen.com/view_correlation?id=292.
3. Tyler Vigen, Spurious Correlations (New York: Hachette Books, 2015).

不要被媒体人骗了

所以，为什么许多人会搞混关联性和因果性？在有些情况下，数据在某种程度上被简化、被夸张、被错误呈现。

还记得我们之前提过的美国有线电视新闻网的一篇文章吧：“研究表明，聪明人用苹果手机。”

如果你单单只看这个标题，你可能会觉得人聪明和用苹果手机之间存在关联。但文章或研究并没有这么说。事实上，美国有线电视新闻网引用的文章根本没用“聪明”一词。[注]这篇文章所依据的白皮书也同样如此。[注]

在这种情况下，媒体所描绘的关联性并非存在于原始研究之中。公平地说，尽管美国有线电视新闻网用了“聪明”一词，但文章并未直言使用苹果手机会让你变聪明。只是有时候关联性会被以一种可能暗示存在因果关系的方式呈现——这是你必须当心的事。

密歇根大学的法学教授J.J.普雷斯科特解释说，大多数人会犯的一个大错误是由于“人们的大脑很容易将因果性和关联性搞混。新的文章时常会碰到一个问题，即事物存在关联性的故事不如事物存在因果性的故事有吸引力。因此，记者甚至学者都会含糊其词，任由读者自己解读”。

还记得在21世纪早期发生的互联网泡沫崩盘吗？当时股市大跌——同时年长者从事劳动的人数上升。科尔博士当时是美国国家经济研究局的老年经济研究计划组成员（需要提一下，这个组织负责给政府提建议，但没有制定政策的权力），她回忆当时有很多媒体报道说劳动率上升是由于人们延迟退休以填补股市的亏空。

“很容易发现这两个现象同时发生，并且看上去好像真的有联系，”科尔对我们说，“但要证明这两者存在因果关系就比较困难。”事实上，她和同事进行了深入的研究，发现“没有证据表明和股市有紧密联系的人群（如持大学学历的工人）和其他工人相比在股灾期间退休得更晚”。他们还发现由于持有股票太多，导致影响到其退休决定的人数很少，以至根本无法在劳动人口变化中体现出来。最终科尔解释道：“尽管报道上说得像那么回事，但在当时改变劳动人口的因素中，股市并没有其他因素重要。”

在接收小数据的时候，记住那些老练的杂志撰稿人、电视制作人、广告文编写人深谙玩弄文字之道，因为他们的工作就是吸引他人注意力。不要被他们骗了。通读全文，思考他们到底说了什么，没说什么。

假设你是一名杂志记者。一天，你看到了一项研究，研究显示吃布朗尼蛋糕和体重增加之间存在关联。编辑让你就此写一篇文章，并把写几个备选标题给她看看。下面就是你的备选标题：

“科学家发现吃布朗尼蛋糕和体重增加之间存在关联。”

“想要减肥，就别吃布朗尼蛋糕。”

“布朗尼蛋糕——最吃不得？”

“体重增加和吃布朗尼蛋糕脱不开关系。”

“吃布朗尼蛋糕会发胖吗？”

“如何通过不吃布朗尼蛋糕来减掉20磅体重？”

上面的标题中，没有一条说吃布朗尼蛋糕导致体重增加。但你可以看到，这些标题暗示两者存在因果关系，只是没有明说而已。

有时候，揭示真相的最佳方法是问问题。如果你看到“体重增加和吃布朗尼蛋糕脱不开关系”这样的标题，你只需问：“体重增加和吃布朗尼蛋糕怎么脱不开关系了？”这个问题的答案会揭示两者之间真正的关系。

---

1. “Infographic: iPhone Usage Rates by State,” Chitika Inc., January 2015, https://chitika.com/insights/2015/iphone-by-state.
2. “Infographic: iPhone Usage Rates by State,” Chitika Inc., January 2015, https://chitika.com/files iPhone_Usage_by_State- Correlation_White_Paper_Chitika Insights.pdf# overlay- context=user/5.

了解大脑的工作模式

那么多人会把关联性和因果性搞混，也许是因为我们都太善于解读数据了。

“人类大脑是一台模式识别的机器。”罗恩·弗里德曼在一次采访中说道。弗里德曼是社会心理学家，专门研究人类动机。他的著作有《最佳工作场所：创造一个非凡的工作场所所需要的艺术与科学》。

“在过去，书籍和搜索引擎还没有发明之前，找出原因和影响对于我们的生存至关重要，”弗里德曼指出，“我们的大脑通过进化，变得善于寻找秩序、预测事情发展。我们无法控制这个过程——我们会处处寻找联系，即使联系并不存在。”

因果关系会让人安心，所以我们想要找到这类关系。

“我们脑中仿佛有这样的程序——揭示事件之间的联系，解读偶然事件，将其作为存在因果关系的证据，”弗里德曼接着说道，“当某件事起因不明的时候，我们会自然而然想要填补这个空白，人为地加上原因。”

如何应对？你无法让自己大脑不要填补空白，但你可以了解大脑如何工作，从而多花一个步骤寻找因果关系的证据。

如果你是波士顿红袜队的死忠粉，然后你觉得只有在你穿自己的幸运T恤的时候他们才能赢球，那你必须明白你正在基于模式制造一个假想的因果关系。[注]

我们十分肯定幸运衫不会扭转乾坤，但这无法改变人们对于事物抱

有期望。我们喜欢理解周围的世界，将关联性变为因果性可以给我们这种理解了世界的感觉，但这并不表示这是解读数据的正确方法。

---

1. For an alternative view exploring the potential psychological influence of such lucky routines, see “Why ‘Magical Thinking’ Works for Some People” by Piercarlo Valdesolo in Scientific America (October 19, 2010), http://www.scientificamerican com/article/ superstitions- can- make- you/.

抛弃先入为主的观念

我们来看一下围绕MMR（麻疹、腮腺炎、风疹）疫苗及其与孤独症关联的争论。或许你听说过明星妈妈珍妮·麦克卡锡谈论过有的家长说自家孩子接种完疫苗以后就发烧，不说话，接着就得了孤独症。注

两者具有关联性——而非因果性。

所以，为什么调查时有1/3的家长认为接种疫苗会导致孤独症？注那篇颇负盛名的声称找到两者联系的期刊文章现已撤回。注一项针对超过95000名儿童的研究发现“接种MMR疫苗并没有增加患孤独症谱系障碍的危险”。注

但如果你孩子在接种完MMR疫苗后，突然变得内向不爱说话，你会怎么办？你会不会怀疑两者之间存在关联？你会不会想要知道为什么你孩子患上了孤独症谱系障碍？

我们觉得你还是要给孩子接种疫苗。科学已经否定了MMR疫苗和孤独症之间存在因果关系的说法，但你可以看到父母的情绪是如何影响他们解读数据的。下次你为容易让人情绪化的事件之间是否存在因果关系争论的时候要记住这一点。

这种寻找答案的想法和实证性偏见有关，该偏见是一种通过解读数据来支撑自己先入为主观念的倾向。怀着实证性偏见的你只是在寻找一个答案——寻找一个特定的答案。

实证性偏见几乎影响着你对待数据的方方面面——从抽样到观察到预测——因此在解读数据的时候，你必须时时留心。在研究关联性和因果性的时候，实证性偏见是有些人忽略遗漏变量的一大原因，因为他们

凭借先入为主的观念，而非基于真实的证据，把两件事时间的关联性当成了因果性。

---

1. Transcripts— CNN Larry King Live, “Jenny McCarthy’s Austism Fight,”CNN website, aired April 2, 2008, http://transcripts.cnn.com/TRAN SCRIPTS/0804/02/lkl.01.html. This is the same interview in which McCarthy says, “I believe that parents’ anecdotal information is science- based information.”
2. “Survey: One Third of American Parents Mistakenly Link Vaccines to Autism,”National Consumers League website, April 2, 2014, http://www.nclnet.org/survey_one_third_of_american_parents_mistakenly_link_vaccines_to_aut
3. A. J. Wakefield, S. H. Murch, A. Anthony, J. Linnell, D. M. Casson, M. Malik, M.Berelowitz, A. P. Dhillon, M. A. Thomson, P. Harvey, A. Valentine, S. E. Davies,and J. A. Walker- Smith, “ Ileal- Lymphoid- Nodular Hyperplasia, Non- Specific Colitis, and Pervasive Developmental Disorder in Children,” Lancet 375, no.9713 (2010) http://www.thelancet.com/journals/lancet/article/ PIIS0140- 6736%2897%2911096-0/abstract. Retracted due to the fact that the investigations were not approved by the local ethics committee and that children were not“consecutively referred.”
4. Anjali Jain, Jaclyn Marshall, Ami Buikema, Tim Bancroft, Jonathan P. Kelly,and Craig J. Newschaffer, “Autism Occurrence by MMR Vaccine Status Among US Children with Older Siblings with and without Autism,” JAMA 313, no. 15 (2015), http://jama.jamanetwork.com/article.aspx?articleid=2275444. Although, if you want to start poking holes, you could start with the fact that the sample set was “privately insured children with older siblings.”

最后但依然重要的一点

哪怕你证明了两件事之间存在因果关系，记住这仅仅说明一件事引发了另一件事，并不能告诉你这个结果的影响或重要性——这两个话题我们会在第五章进行讨论。

如果你只是想要知道是否X引起了Y，那证明两者存在因果关系就足够了。但是如果你想要知道X（然后就是Y）是如何影响你日常生活的，那你就要多问几个问题了。

所以，该如何处理因果关系，如何切实证明两者关系是因果关系而不是简单的相互关联？

我们发现我们所提的很多建议都在提醒你应该注意哪些陷阱，哪些事不该做。这本书并不是统计学教材。很不幸，我们没有足够的版面来教你如何做一项完美的统计学分析，或确定两个事物到底存在怎样的关联。但没事，因为我们的目标就是想要帮助你通过明白关联性和因果性的区别，了解人们为什么会将这两者搞混来让你在做决定的时候更为明智——不重蹈前人的覆辙。

如何成熟地应对关联性和因果性

现在，对关联性和因果性之间的差别有了较好的理解，我们在接收有关统计学关系的数据的时候，还有几点要谨记于心：

1.问你自己，**这篇新发表的文章或研究展示了什么。**文章中的确用了“因果”关系这个字眼了吗？不少时候，标题或文章中可能会暗示因果关系，但如果你深入研究，会发现大多数实际的研究只是在讨论某种关联性。

2.在理解统计学分析结果的时候，退一步，用**常识思考一下**——这样的关系从直觉看来对不对？为什么吃烤奶酪可以提高性生活质量？聪明人用苹果手机这种说法能不能说得通？尽管统计学常常可以得出出人意料的结论，但不要把一切孤立起来看待，不要抛弃自己的直觉。

3.如果你看到了两者之间的关系，问自己：**会不会有其他的因素导致了我所观察到的结论？**在理解两者关系的时候，是不是有其他确实非常重要的遗漏变量。

4.**时刻当心反向因果关系。**找到统计学上的关联性并不能表示事情就是按照那个顺序排列的。聪明人会晚睡吗？或者人们晚睡正是因为他们聪明？不要小看反馈循环——X影响了Y，Y同时又影响了X（如聪明人晚睡，而晚睡又给了人们更多时间让自己变得聪明，而人们变聪明了，又可以晚睡了……）

5.最后，对于科学家来说，**证明因果关系也是一件非常需要技术含量的事。**对于那些据称是因果关系，尤其可能存在潜在的遗漏变量的情况，要特别留心。

第五章 眼见真的为实吗？我们信仰统计学

哪怕在华盛顿这个到处是宏伟纪念碑的城市里，越战阵亡战士纪念碑依然与众不同。几十块黑色花岗岩墙面上刻着在越战中阵亡和失踪人员的名字——总数超过58000个。

但由于一个据称是随机的抽签程序使得有些美国军人死去，而另一些人活了下来。[注]

1969年，美国在越南有超过50万军队。美国兵役注册部门开始了二战以来第一次征兵抽签，以决定下一年哪些人要继续服役。1969年12月1日将进行一项随机抽签，约有85万“符合抽签条件”的年轻人将会根据生日的不同分配到不同的抽签数字。

抽签是这么运作的：工作人员将366个蓝色塑料胶囊放到盒子中，每个胶囊中都内含一个从1月1日到12月31日之间的日期[注]。再将盒中的所有胶囊倒进一个大型玻璃容器里，工作人员随后将胶囊一一从玻璃容器中取出。

如果第一个胶囊中写着你的生日（9月14日），那你就被分配到1号；如果你的生日出现在第二个胶囊中（4月24日），那你就被分配到2号；以此类推。编号越往前，你就越早服役。据估算如果你的编号排在前1/3，那你可能就要准备在不久的将来前往新兵训练营了，接着开赴越南。

从理论上看，兵役注册部门选人是随机的。每个人被第一个抽到的概率相同，但事实并非如此。

1970年《纽约时报》上有一篇文章的标题为——《统计学家指出征兵抽签并非随机》。[注]该文章指出，12月出生的人和1月出生的人相比，会更容易在抽签中编号靠前。事实上，如《纽约时报》上的一幅图表显示，生日从1月到5月间的人在抽签中的编号平均数大于200，再往后编号逐月变小（除10月份外），最终减到12月的122。

在本章中，我们准备探寻科学工作者在判断某种统计学效应是否为随机上使用的不同方法。

根据接受《纽约时报》采访的统计学家的说法，在这种情形下，随机出现图表上结果的概率为1/50000。

那如果这并非随机，那又该如何解释这种倾向？

从统计学角度，在某种特殊的情况下，比较靠后的月份的编号平均数会较小（较早被抽到）。那是在什么情况之下呢？即靠后月份的胶囊处于玻璃容器上层，而非与其他胶囊随机混合。

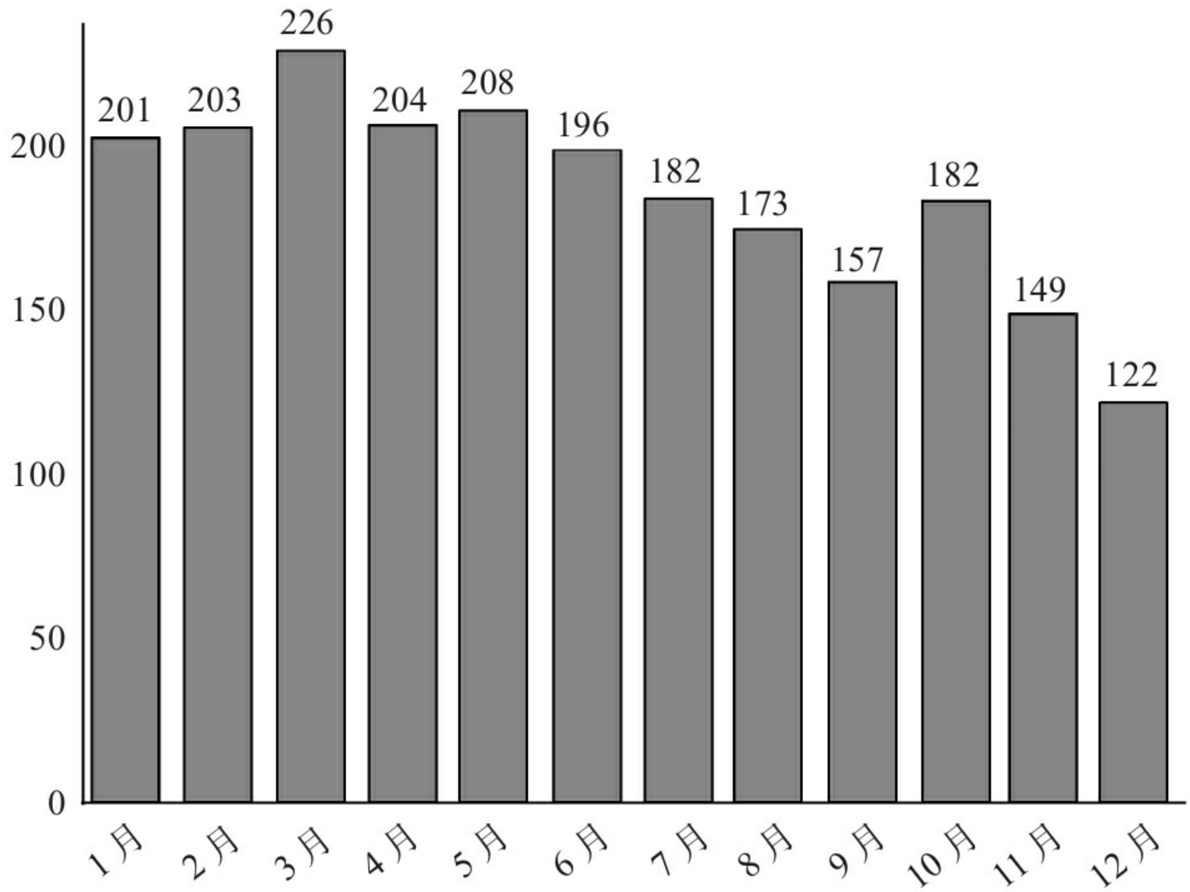


图5－1 不同月份出生的人所获得的抽签编号平均值

兵役注册部门的公共信息部主任对《纽约时报》解释了胶囊是如何混合的。首先，工作人员将1月份的31颗胶囊装入容器。“1月份的胶囊被放入了一个方形大木箱，接着工作人员用硬纸板将胶囊推到箱子一侧，将箱子另一侧空出来。”注接着，工作人员将2月份的胶囊倒入箱子空着的一侧，然后再用硬纸板将其推向一侧。投入每个月份的胶囊都会经过这样的流程。这样，1月份的胶囊就会和其他月份的胶囊混合11次（2月份胶囊也一样，因为1月2月的胶囊是第一组进行混合的——《纽约时报》没有搞清楚这个事实），接着3月份的胶囊和其他胶囊混合了10次，4月份的9次，以此类推。

12月份的胶囊仅和其他月份胶囊混合了一次。

兵役注册部门以这个方式抽签也有他们自己的考虑——也许早在20世纪40年代那次工作人员用木勺搅动胶囊的时候把胶囊弄破的事件也是原因之一。（根据《科学》杂志的说法，那个木勺由“费城独立纪念厅椽条的木料”制成）。注

至少从历史记录上看，并不清楚胶囊是通过何种方式倒进玻璃容器中的，但无疑你可以看出以逐月方式加入并混合胶囊可能导致不同的结果。试问如果将混合胶囊的木箱掉转方向，从另一个口将胶囊倒入玻璃容器，最终的结果会不会截然相反？注

在这个例子中，抽签号码本该是“随机”抽取的；因此，如果号码的确被抽到的概率相同，就应该有一个基准结果。从统计学角度，如果抽签结果和基准结果不同，那就可以明显地得出由生日编号组成的抽签结果很可能并非出自一个纯随机的过程。（尽管使用生日——如果方式正确的话——可以作为一种随机的方式来解决问题。）

作为一个成熟的小数据接收者，你必须能够辨别从数据中找到的关系真实性有多大。你在数据中找到的关系是随机产生的，还是另有奥妙？你有多少把握可以肯定你所看到的是真实的——你对其影响衡量的准确度和精确度有多少？我们在本章中将会进一步探讨这几个问题。

---

1. Randomization is a powerful statistical tool for eliminating selection bias, and Esther Duflo and others have written extensively about randomized controlled trials and related topics. For more detail, “Using Randomization in Development Economics Research: A Toolkit,” © 2006 by Esther Duflo, Rachel Glennester, and Michael Kremer, National Bureau of Economic Research, 2006,http://www.nber.org/papers/t0333.pdf.

2. February 29 was included in the dates to account for men born in leap years.

3. David E. Rosenbaum, “Statisticians Charge Draft Lottery Was Not Random,”New York Times, January 4, 1970, http://frewm.wikispaces.com/file/vie/nytimes.pdf. If you look at the original article, you’ll see that the average number for July was reported as 180. Based on our calculations, it was actually 182.

4. Rosenbaum, “Statisticians Charge Draft Lottery Was Not Random.”

5. Stephen E. Fienberg, “Randomization and Social Affairs: The 1970 Draft Lottery,” Science 171 (1971): 255, http://conallboyle.com/lottery/05USmil_draft1970.pdf.

6. We might also be able to learn something about the process by studying the variance— the amount by which values vary— within each month.

民意调查

2015年夏天，《彭博社政治版》针对共和党人做了一项投票，询问他们心中美国总统的最佳人选。

10%的人选择了杰布·布什，8%的人选择了斯科特·沃克，这个结果显示布什仅领先沃克2%。[注]

也许是这样。因为当你观察数据，你会发现有可能布什实际上落后6%，也可能布什领先10%。

这个结论是如何得出来的？彭博社投票的误差范围为±4.4%。[注]

每次调查中，总会有些“摇摆不定的投票者”，他们会给调查结果带来随机性。误差范围便是衡量这类投票中不确定因素的常用统计学手段，是回答“你真的确定吗”这个问题的一种方式。

很多人曲解了误差范围，认为其表明了竞选人的支持程度肯定在这个范围之内。这种定义不太正确。

误差范围到底指什么，根据彭博社的说法，指“如果一项调查重复使用相同的问题和相同的操作方式，调查结果的偏差程度20次中会有19次在±4.4%范围之内”。[注]

因此，回到彭博社的投票——为什么我们无法得出100%相同的结果？为什么误差范围会这么大——增加4.4%或减少4.4%？

因为投票便是抽样。在上面的情形中，取了500人的样本。样本并非总体——在上次总统选举中，有超过6000万人把票投给了共和党。在

这种情况下，给如此大的数据总体抽样，必须要面对抽样的误差，这事我们在本章会详细讨论。[注]

这就说明布什可能——或可能没有——在选举的时候真的领先沃克。在存在误差范围的情况下，根据现有数据很难下结论。我们对此次投票结果的最佳解释是布什领先于沃克。但考虑到误差范围的大小，我们无法确定真实情况是否如此。

最后，给你呈上今天最有趣的一个事实：根据美国国家经济研究局的说法，“福克斯新闻频道的报道对1996—2000年的总统选举投票产生的影响不大，但具有显著性差异”。[注]

---

1. John McCormick, “Donald Trump Dominates Republican Field in Pre- Debate Bloomberg Poll,” Bloomberg, August 4, 2015, http://www.bloomberg.com/politics/articles/2015-08-04/donald- trump- dominates- republican- fiel-in- pre- debate- bloomberg- poll.
2. With a margin of error of ±4.4 percent, the researchers would expect that Jeb would not be more than 4.4 percent higher or lower than 10 percent (approximately 6 to 14 percent), and that Walker would not be more than 4.4 percent higher or lower than 8 percent (roughly 4 to 12 percent).
3. “Bloomberg Politics National Poll,” Bloomberg, August 4, 2015, http://images.businessweek.com/cms/2015-08-04/8302475320_tue.pdf.
4. Of course, there may be other factors that affect sampling error— such as whether you’re polling all voters versus likely voters, how you’re conducting the poll (some pollsters may only call land lines, not cell phones), etc.
5. Les Picker, “Media Bias and Voting,” National Bureau of Economic Research,http://www.nber.org/digest/oct06/w12169.html.

二手烟是否会致癌？

在美国国家癌症研究所的网站上，有一个页面专门用来共享有关二手烟和癌症的信息。[注]页面上第4个问题看起来十分直截了当：“吸二手烟会致癌吗？”

答案：“会。”

国家癌症研究所声称提供“科学、权威、最新”的数据”[注]，并列出许多文献，这些文献将二手烟归于癌症诱因。

该网站指出“吸入二手烟会导致不吸烟的成人罹患癌症”。该网站指出和吸烟者共同生活“会使不吸烟人群患上癌症的概率增加20%~30%”。该网站甚至引证了美国每年由于二手烟而死亡的不吸烟者人数——约3000人。[注]

所以可以想象当我们读到《国家癌症研究所期刊》上一篇名为《被动吸烟和癌症不存在明显关联》[注]的文章时有多惊讶。

啊，现在该如何是好？我们手头有国家癌症研究所数十年的研究成果，声称二手烟会引发癌症；同时也有这个研究，声称两者不存在明显关联。

你该如何解读这些数据？

你可以花去数天（也可能是数星期或数月）时间通读国家癌症研究所的所有参考文献，外加一份美国卫生局局长所撰写的727页的报告，试着自己来寻找答案。

你可以听取针对这项最新研究的资深研究员的说法：“我们认为文章想要传递的信息便是基于分析，我们无法得出危险是什么，甚至无法得出是否存在危险。”

或者你也可以停一下，并思考研究员和科学家如何判断他们的研究结果是否可信。

因此，作为一个成熟的数据接收者，应该怎么办？

---

1. “Secondhand Smoke and Cancer,” NIH National Cancer Institute website, reviewed January 12, 2011, http://www.cancer.gov/ about- cancer/ causes- prevention/risk/tobacco/ second- hand- smoke- fact- sheet#r5.
2. “About This Website,” NIH National Cancer Institute website, posted April 10,2015, http://www.cancer.gov/ about- website.
3. This number appears to be an estimate from an EPA report: “Approximately 3,000 lung cancer deaths per year among nonsmokers (never- smokers and former smokers) of both sexes are estimated to be attributable to ETS in the United States. While there are statistical and modeling uncertainties in this estimate,and the true number may be higher or lower, the assumptions used in this analysis would tend to underestimate the actual population risk. The overall confidence in this estimate is medium to high.” U.S. EPA, “Respiratory Health Effects of Passive Smoking (Also Known As Exposure to Secondhand Smoke or Environmental Tobacco Smoke, ETS?),” U.S. Environmental Protection Agency, Office of Research and Development, Office of Health and Envirmental Assessment, Washington, D.C., EPA/600/6-90/006F, 1992.
4. Judy Peres, “No Clear Link Between Passive Smoking and Lung Cancer,” Journal of the National Cancer Institute 105 (2013): 1844– 1846, http://jnci.oxford journals.org/content/early/2013/12/05/jnci.djt365.full.

重要的事

在电影《感谢你抽烟》中，艾伦·艾克哈特扮演的角色（烟草公司发言人）对他儿子说：“只要以正确的方式辩论，你就永远不会是错的。”注

上面的话尽管只是好莱坞讽刺片一个说客的台词，但依然很有趣，在我们谈论统计学中显著性差异的时候，需要记得这句话。很多人觉得显著性差异便是谈论数据的“正确”方法。

显著性差异是科学家和学者设定的客观标准，用来确定“从统计学角度来看”某个特定的关系是否存在于数据中。科学家通过检验显著性差异以辨别观察到的效应是否在数据中有所体现（存在很大可能性），或仅仅出于巧合。有一点很重要，即找到显著性差异并不能告诉我们某种关系是关联性的还是因果性的，同时也不能告诉我们是否存在遗漏变量影响着最终结果。

显著性差异指某事真实的可能性。可以用来衡量我们观察到的结果有多少概率是真实的（而非出于巧合），所以显著性差异通常会用P值来衡量。P在这里指可能性。如果你接受使用P值作为显著性差异的衡量标准，那P值越小，你观察到的结果仅仅出于巧合的概率就越低。注

衡量显著性差异时，普遍为人所接受的P值为小于0.05（相当于5%可能性）。这个数值得到广泛应用的时间可以追溯到20世纪20年代，推广者为数学家罗纳德·费希尔，他的研究方向为肥料对作物产生的影响。注

我们并非要在这里辩论0.05的P值是否是衡量显著性差异的合适标

准，我们甚至不想讨论P值本身是否是衡量显著性差异的合适方法。[注]

相反，我们在这里是想告诉你P值——包括0.05这个值——在很多场合下就是标准。

这就是为何这两者对你来说很重要。

因为当你看到一篇有关最新科学发现的文章，很可能就因为其P值低于0.05，才被科学界接受，并被媒体所报道。

这可能看起来比较武断，但就像布法罗大学副教授德瑞克·丹尼尔博士对我们说的："有一条标准线可以让我们保持客观。如果没有这一条线，那我们就会在对自己有利时对P值0.06大肆宣扬，而故意忽略不利于自己的P值0.04。"[注]

---

1. Thank You for Smoking, Fox Searchlight Pictures website, accessed August 9,2015, http://www.foxsearchlight.com/thankyouforsmoking/.
2. Ultimately, a statistical analysis is trying to determine if there is enough evidence to reject a null hypothesis. As Jesse Farmer wrote on the 20bits blog, "Don't read anything into the fact that it's called the 'null' hypothesis— it's just the hypothesis we're trying to test." (Jesse Farmer, "Hypothesis Testing: The Basics," 20bits blog, April 22, 2009, http://20bits.com/article/hypothesis- testing- the- basics).Typically, a null hypothesis is usually one where there is no relationship between two variables. In other words, your data can't tell you one way or the other whether or not there is a relationship— it could simply be due to random chance.When you reject the null hypothesis, it means that there is an association that is likely not due to chance.
3. That said, even Fisher appears to have had some flexibility on the .05, saing. "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred(the 1 per cent point)." Ronald A. Fisher, "The Arrangement of Field Experiments," Journal of the Ministry of Agriculture of Great Britain 33 (1926): 503–513, accessed August 9, 2015, https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15191/1/48.pdf.
4. Regina Nuzzo, "Scientific Method: Statistical Errors,"Nature website, February 12, 2014, http://www.nature.com/news/ scientific method- statistical- errors-1.14700.
5. As Daniels noted, this approach is not what hypothesis testing is all about. In our opinion,

it’s also important to keep in mind that “statistically significant,” as many people understand it (p<.05), does not mean the findings are 100 percent certain. The p-value measures probability, which means that some results may still be more probable than others. For example, p-values of .001 and .049 are both statistically significant (by this definition), but a lower-value means that these results are more likely to be true. (To be technically precise, it actually means that finding a lack of a relationship between two variables is more likely to be false.)

深呼吸

现在我们回看那份二手烟的研究，看一下这份研究报告到底说了什么——被动吸烟“在统计学上并没有显著增加肺癌危险”。

但是，正如那个“不存在明显关联”的研究者指出，他们的一部分结论是基于相对较小的人群得出的。在所研究的76000人当中，只有152人从未吸烟而且得了肺癌。这有问题吗？也许有问题。

是否这项研究仅将女性作为研究对象？有可能。

这项研究仅研究了暴露在二手烟环境中的年数，而非强度（一年多少包），这又说明了什么呢？不管怎样，一个不抽烟的妻子，其丈夫在走廊抽烟；另一个不抽烟的妻子，其丈夫在室内抽烟，两人比较的结果应该会有所不同。

的确，很可能这一切因素都会产生影响，但关键在于这项研究的大部分无法表明吸二手烟和癌症存在统计学上的关系。[注]这并不表示两者就没关系了。而是表明了研究人员无法通过研究将两者从纯随机事件中区别开来。

最后，我们谈论显著性差异时，可能会碰到一些一定要将P值控制在0.05的统计学原理（包括标准差和其他衡量手段的理论），这些原理设定了一个严格的界限，让一些没有达到这个值的人觉得恐慌。事实上，有人在博客上发表了509个经过同行评审的期刊上有关接近（但没有达到）显著性差异的“戏谑说法”。[注]其中有些说法我们很喜欢，如：

“与显著性擦身而过”。

“与传统意义上的显著性调了一次情”。

“就差那么一点就显著了”。

“类似显著”。

“十分接近显著”（当时的情形下，P值为0.05009）。

“在显著的悬崖边艰难迈步”。

就着以上的表达，我们继续。

---

1. They did find one group of women— those who lived with a smoker for at least 30 years— that did show a higher risk for lung cancer, but even that finding was,according to the researchers, “of only borderline statistical significance.

2. Matthew Hankins, “Still Not Significant,”Probable Error blog, April 21, 2013,https://mchankins.wordpress.com/2013/04/21/ still- not- significan-2/.

抽样大小事关重大

《纽约客》等杂志曾经刊登过一项研究。在那项研究中，研究人员做了一系列实验，以测验人们是否在饥饿的时候更倾向于购物。注

他们是如何研究这个现象的呢？在其中一个实验中，他们调查了81位从百货商店出来的顾客。现在，我们站在商店出口处，看到店里的顾客似乎不止81位，因此我们很确定这个数字并非全世界顾客的总数。这个结论引出了确定显著性差异的关键因素——抽样。

回忆一下第二章的内容，抽样可以让我们从数据总体中估算结果。例如，问每一个美国人他们在总统竞选的时候想投谁的票是不可能实现的事。如果那样，你就得打很长时间电话了。因此，民意调查分析人员会观察总体的样本。

但是抽样并非无懈可击。

我们在第二章里讲过抽样面临的一个问题，就是你可能想解决某个问题，却观察了错误的数据。还记得“挑战者号”空难吗？调查组观察了过去执行发射任务时气温的样本，但只看了出现O型环事故的样本——他们应该看的是所有发射任务的样本。

但现在，我们假定你观察的是正确数据，试想一下另一个重要的问题：样本是否从统计学角度告诉你有关总体的信息？换句话说，你有多少把握可以肯定样本正确地体现了其余数据——数据总体？

简单来讲，观察样本总会存在一定的不确定性。用统计学术语说，这种现象叫作样本误差。样本误差体现了你的样本有着多大的不确定性。样本误差之所以存在，是因为并非所有的样本都是相同的。哪怕你

从同一个数据总体取了两个相同大小的样本，依然可能得到不同的结果。试想你从一个装满M&M巧克力豆的碗里随机取出5颗巧克力豆，然后将其放回碗中，接着再取5颗。两次取到相同颜色组合的巧克力豆的可能性有多大？

研究人员会花大量时间检验其研究的功效——简单来讲即抽样的大小和其推算统计结果能力之间的程度关系。你作为小数据的接收者，只要注意样本大小（一个具体研究结果所基于的数据点量）对研究结果是否存在显著性差异起着至关重要的影响。

---

1. Alison Jing Xu, Norbert Schwarz, and Robert S. Wyer, Jr., “Hunger Promotes Acquisition of Nonfood Objects,” Proceedings of the National Academy of Sciences 112, no. 9 (March 3, 2015), http://www.pnas.org/content/112/9/2688.abstract;read the New Yorker overview here: Nicola Twilley, “The Good, the Bad, and the Hangry,” New Yorker website, March 10, 2015, http://www.newyorker.com/tech/elements/ hunger- good- bad- hangry.

你有多大把握？

还记得误差范围是我们在投票时表达统计置信度的一个手段吗？在衡量科学研究结果的统计可信度方面，还有一个类似（但稍有不同）的工具，名为置信区间。该度量标准通常以一个区间值的形式出现，而不像误差范围一样用“±”表示。但这两者的功能相同，都显示了你可能会估计（当然，假设你的样本是随机选取的）的一个区间值。

和误差范围一样，置信区间越大，区间内包含（整个数据总体中的）真实值的可能性越大。从这个角度设想：如果布什的得票率为10%，那很可能他真实的支持率为5%到15%之间的某个值。从统计学上讲，他的真实支持率不太可能落于一个较窄的区间，如8%到12%。

于是便涉及了置信水平——该词用来表示我们正确度量数据的可信度有多高。置信水平通常以百分数形式出现，表示样本包含真实值的概率有多大。

置信区间和置信水平在科学论文和科学研究中应用广泛，而在针对科学论文和研究的媒体报道中所见甚少。这存在问题，因为一旦离开了置信区间和置信水平，就无法知道事情的全部真相。

比如，美国国家公共电台有一篇文章说每天喝含糖饮料的人群“10年后患上（肥胖症）的概率会上升18%”。[注]根据研究报告的说法，这是真的。但如果你读了原始材料，你会发现18%这个值“置信区间为8.8%到28%，而置信水平为95%”。[注]所以的确有可能患病概率会上升18%，但更加准确的表达方式应该是患病概率的幅度在8.8%~28%之间。

置信区间要求数据接收者对基本数据有细致的理解，而单看媒体报

道很可能做不到这一点。我们并非针对美国国家公共电台、美国有线电视新闻网或者任何一家媒体——尽管它们深深伤害了我们这些热爱数据的人，大多数人并不在意所有的细枝末节。这没关系，你不需要了解自己接触到的每一个研究的置信区间，但你需要知道置信区间是存在的、置信区间的定义，以及置信区间是如何影响你每天接收到的数据的。抑或，用唐纳德·拉姆斯菲尔德的话说，你要能够分辨那些“你不懂，也不知道其存在的东西”和“你不懂，但知道其存在的东西”。[注]

当然，媒体对数据的解读和本来结论就可能不正确的原始科学研究是两回事。约翰·安尼迪斯在一篇题为“为什么绝大多数已发表的研究结果是错误的”的文章中写道：“有一个越来越令人担心的问题，即已发表的研究报告中，有大多数，甚至绝大多数报告所得出的结论是错误的。”[注]

《科学新闻》有文章指出：“如果你还信任科普读物的话，你就不该相信科普读物文章中所写的东西。”[注]

我们并不确定安尼迪斯声称的“大多数”研究结论是错误的这一说法是否属实，但我们的确看到过几十个——也可能是几百个——公开发表的研究在显著性差异方面做得令人担忧。现已出版的研究中，数据无法支持结论的情况并不罕见。

---

1. Allison Aubrey, “Even if You’re Lean, 1 Soda Per Day Ups Your Risk of Type 2 Diabetes,” NPR website, July 23, 2015, http://www.npr.org/sections/thesalt/2015/07/23/425635400/even-if- youre- lean-1- soda- per- day- ups- your- risk-of-diabetes.

2. Fumiaki Imamura, Laura O’Connor, Zheng Ye, Jaako Mursu, Yasuaki Hayashino, Shilpa N. Bhupathiraju, and Nita G. Forouhi, “Consumption of Sugar Sweetened Beverages, Artifically Sweetened Beverages, and Fruit Juice and Incidence of Type 2 Diabetes: Systematic Review, Meta- Analysis, and Estimation of Population Attributable Fraction,” BMJ 351(2015), doi: http://dx.doi.org/10.1136/bmj.h3576.

3. On February 12, 2002, Rumsfeld (former U.S. secretary of defense) appeared at a U.S.

Department of Defense briefing and said: “There are known knowns;there are things we know we know. We also know there are known unknowns;that is to say, we know there are some things we do not know. But there are also unknown unknowns— the ones we don’t know we don’t know.” On Rumsfeld’s website (http://papers.rumsfeld.com/about/page/ authors- note), he says he “first heard a variant of the phrase ‘known unknowns’ in a discussion with former NASA administrator William R. Graham, when we served together on the Ballistic Missile Threat Commission in the late 1990s.”

4. John P. A. Ioannidis, “Why Most Published Research Findings Are False,” Public Library of Science Medicine 2, no. 8 (August 30, 2005): doi:10.1371/journal.pmed.0020124.

5. Tom Siegfried, “Odds Are, It’s Wrong,” ScienceNews 177, no. 7 (March 27,2010): 26, https://www.sciencenews.org/article/ odds- are- its- wrong.

泄露秘密

如果你读了《今日医学新闻》，你可能会看到这样一个新闻标题：“研究显示喝咖啡可以降低患上子宫内膜癌的风险。”㊟

该文章解释了什么是子宫内膜癌（你可能将其当成子宫癌），并告诉读者有一项研究以及这项研究的发现，即每天喝大约4杯咖啡能够降低女性患子宫内膜癌的风险。

接着，我们来看一下文章的第6段：“受访女性完成了有关其饮食的调查问卷，研究人员评估了其摄入的84种食物及营养元素，以及患上子宫内膜癌的风险。”㊟

因此，研究人员研究了84种不同的东西，并且找到了其中一种食物能够降低患癌风险。㊟

你应该记得，显著性差异通常基于找到结果的可能性而得出。因此，你研究的对象越多，就越可能“发现”显著性差异。统计学家称其为多重比较问题。如果你检验一个关系，并且以95%可能性为标准判断显著性差异，那将会有98.7%的概率在84次测试中纯粹随机找到至少一个正面结果。

咖啡是世界上人们研究最多的食物之一。事实上，Vocative.com网站上有一篇名为《我们放弃了，我们一起来说咖啡是万能灵药吧》的文章，文章中指出有“约2000篇论文将咖啡称为某种预防疾病的药物”。㊟

足足2000篇论文在写一种含有1000多种化学成分的饮料。

说实话，如果有研究找不到喝咖啡的任何好处，我们反而会觉得惊

讶。

经济学家J.J.普利克特给过一个建议，我们应该记住：缺少效应的证据（对于统计学发烧友来说，即无法拒绝零值）不同于有证据证明缺少效应（接受零值）。换句话说，单单因为无法证明某事发生，并不意味着某事就没有发生。哪怕你没有听见（或看见）树林里有树倒下，树依然有可能倒下了。

---

1. Honor Whiteman, “Study Links Coffee Intake with Reduced Risk of Endometrial Cancer,” Medical News Today website, February 6, 2015, accessed August 13, 2015, http://www.medicalnewstoday.com/articles/288988.php.
2. Whiteman, “Study Links Coffee Intake.”
3. We thank Susan Dynarski for pointing this example out to us.
4. Joshua A. Krisch, “We Give Up. Let’s Just Say Coffee Cures Everything,”Vocativ website, January 26, 2015, http://www.vocativ.com/culture/ health- culture/ coffee- cures- cancer/.

如何面对截然相反的结论

当今世界，很容易在公开发表的研究报告中找到截然相反的结论。二手烟会——不会——导致肺癌。葡萄酒也一样。还有牛奶、蛋、番茄、马铃薯、咖啡，甚至玉米……根据不同的研究报告，会导致（并且预防！）癌症的食物名单数都数不到头。

事实上，有两个研究人员决定对这个现象展开研究，他们的研究报告题为《我们吃的所有东西是否都和癌症有联系？系统性菜谱研究》。

他们首先从菜谱中选取了50种食材，并对此展开研究，后发现50种食材中有40种“有文章表明会致癌”。[注]

但是有些文章说食用这些食物会提高患上癌症的风险，而另一些文章指出同样的食物事实上会帮助预防癌症。

所以，你该如何决定正餐吃些什么呢？

显著性差异如果使用正确的话，会是一个非常有用的工具。经济学家贾斯汀·沃尔弗斯是布鲁金斯学会会员、密歇根大学教授，长期担任《纽约时报》经济版的专栏作家。我们问他，当碰到这么多互相矛盾的研究报告时，该怎么做。他解释道：“通常来说，单凭一项研究是无法推翻现存的文献结论的，必须结合该领域所有其他研究一起来看。原因就在于有一项存在显著性差异的研究表明X能够预防癌症，并不能表明不存在另一项研究表明X不能预防癌症。”

---

1. Jonathan D. Schoenfeld and John P. A. Ioannidis, “Is Everything We Eat Associated with Cancer? A Systematic Cookbook Review,” American Journal of Clinical Nutrition 97 no. 1 (2013): 127–134, doi: 10.3945/ajcn.112.047142.

效果显著

假设你在一家制药公司工作，你公司制造一种能够延长患有ALS（卢伽雷氏病）病人生命的药品。你经过研究，确认该药物在统计学上的确有显著效果，该药品是有疗效的。

现在，关键问题是该药品的疗效有多好？可以延长病人一天生命，还是一个月生命，还是一年生命？

用统计学术语说，这称为效应尺度。从本质上说，尺度即效应量，是判断结果是否有意义的一种方式。没有尺度，就难以了解某事有多重要。思考下面两个陈述：

喝咖啡能够降低患上癌症的风险。

喝咖啡能够将患上癌症的风险降低18%。

18%即尺度——（根据这项研究）咖啡所能引起的变化。

效应量和是否存在显著性差异不同。某事存在显著性差异并不意味着它的效应是巨大的抑或是微小的。

需要记住，效应尺度会根据关系的不同产生变化。比如，你可能会同意一个平日里常听到的说法：喝水对你有好处。但喝水对你有多大好处取决于你喝多少水。为FiveThirtyEight网站撰文的欧斯特博士（从《美国流行病学期刊》中引用了一篇研究报告）发现"喝多一点水能够降低死亡危险"。[注]不论男女，每天喝超过三杯水是可以降低死亡危险的，但如果你喝了过量的水，则会致命。

在你看到结果的时候，想一下效应尺度。比如，如果你读到了一项

研究指出“咖啡可以预防子宫内膜癌”，了解一下喝咖啡的量很重要。你只需喝一杯就能获益，还是你必须喝4杯才行？

---

1. Emily Oster, “You Don’t Need 8 Glasses of Water a Day,” FiveThirtyEight Science website, posted September 30, 2014, http://fivethirtyeight.com/features you- dont- need-8-glasses-of-water-a-day/.

这个研究对我的生活重要吗？

就算你手头的研究存在显著性差异，就算你手头的研究效应很大，只要你讲的是在日常生活中接收数据，那你能问的最关键问题就是："这个研究将对我的生活产生怎样的影响？"拿我们所说的咖啡例子来看。多喝一杯咖啡对降低癌症风险有多大"好处"？多喝咖啡能把癌症风险降低0.00001%还是10%？

回答这个问题能够凸显统计影响和经济影响的不同。

统计影响就是简单地说明是的，存在某种大小未知的关系。

但大多数人不会基于统计学的关联性做决定，而会观察经济影响——这个决定会花费我们多少时间、金钱、健康或其他资源。

我们询问欧斯特博士（她决定在孕期吃腌制肉类，尽管吃这个食物会有一定的概率感染李斯特菌）有关效应尺度、统计影响和经济影响的对比，以及研究人员和大众的关注点区别等问题。

"我觉得应该这么讲，"她说，"理论上的危险和足以引起人们注意的危险之间是有差别的。在我的例子里，涉及了李斯特菌——的确存在腌制肉类中含有李斯特菌的情况，但也有案例表明李斯特菌存在于菠菜、哈密瓜、冰淇淋中。所以你的确可以通过不吃腌制肉类在极微小的程度上降低感染风险——因此可以理解有些人会做出这样的选择，但降低危险的程度很小。"

显著性差异起了作用。

但在许多情形下，日常生活中，效应尺度和经济影响起了最大的影

响。

等等，还有呢

显著性差异并不能支配一切。哪怕你已经得出了显著性差异，依然需要留心我们在本书中提到的其他因素（遗漏变量、离群值等），除此之外还有其他的一系列偏差（证实偏差、选择偏差等）。

请思考：

•在子宫内膜癌的研究中，根据WebMD（美国最大的医疗健康服务网站）的说法，研究人员“找到了喝咖啡与降低子宫内膜癌风险之间的联系，但这个联系并非因果联系”。注也就是说，这两者之间存在关联性，而非因果性。

•根据《PLoS生物学》杂志的说法，“P值黑客”（得名于P值）一词指研究人员“收集、选择数据或统计学分析材料，直到非显著结果变成显著为止”。注这个方式与“有目的地选择”类似，打个比方，就好比研究人员抓到什么就往墙上扔，直到有东西粘在上面为止（尽管有可能有些科学家真的会往墙上扔东西，直到粘上墙为止……）。

•《纽约客》上有一篇精彩的文章（除此之外还有其他不精彩的文章吗？）讲到发表偏倚可能会导致让显著性差异效应逐渐降低的“降低效应”。为什么？一位统计学家发现“已发表的带有显著性差异的心理学研究中，有97%得出了想要得出的结果”，这可能会让未来的研究难以重复得出这些结论。注

•《流行病学与社区健康期刊》上有一篇文章指出没有证据表明在英格兰与威尔士减少街道夜晚光照会导致交通事故和犯罪率上升。但文章作者正直地承认文章可能存在选择偏差——他们没能从大约1/3的当地有关部门获得数据，并说：“当地有关部门可能不愿合作，因为其认

为，或知晓当地交通事故和犯罪率上升是减少街道夜晚光照引起的。”
注

仅仅知道某事存在显著性差异，并不表示就万事大吉了。

---

1. Kathleen Doheny, “Coffee May Lower Endometrial Cancer Risk,” WebMD website, February 6, 2015, http://www.webmd.com/cancer/news/20150206/ coffee- linked-to- possible- lower- endometrial- cancer- risk.
2. Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions, “The Extent and Consequences of P-Hacking in Science,” Public Library of Science (PLoS) Biology 13, no. 3 (2015): e1002106, doi: 10.1371/journal.pbio.1002106.
3. Jonah Lehrer, “The Truth Wears Off” New Yorker website, December 13, 2010,http://www.newyorker.com/magazine/2010/12/13/ the- truth- wears- off.
4. Rebecca Steinbach, Chloe Perkins, Lisa Tompson, Shane Johnson, Ben Armstrong, Judith Green, Chris Grundy, Paul Wilkinson, and Phil Edwards, “The Effect of Reduced Street Lighting on Road Casualties and Crime in England and Wales: Controlled Interrupted Time Series Analysis,” Journal of Epidemiology Community Health (June 3, 2015), doi: 10.1 136/jech- 2015- 206012, http://jech.bmj.com/content/early/2015/07/08/jech- 2015- 206012.full.pdf+html.

了解自己所看到的是否重要，从而成为成熟的数据接收者

仅因为自己被数据包围，并不表示你就要使用这些数据在生活中做决定。下面有5件你可以立即着手去做的事，以此了解自己所看到的数据是否真的都有用。

1.确定所看到的结果是否随机出现。作为一个成熟的数据接收者，常常需要在观察到的结果中排除那些随机出现的。你也许接触了5个喜欢吃汉堡的男性，但以这个样本来判断所有男性中有百分之几喜欢吃汉堡，这个样本可能就太小了。也许你遇到的这5个人，是世界上唯一喜欢吃汉堡的人群。在许多情况下，判断结果是否随机需要一个基线，以此来比较你所得出的结果。

2.要了解，**许多研究发现其实是基于或然性的。**一个具有“显著性差异”的研究发现，也仅仅表明结论有95%的可能性落在置信区间内。仔细研究P值可以让你了解结论有多少可能性并非仅仅是随机出现的——很多情况下我们会得出可以确信的结论，但要记住，我们只是在衡量或然性而已。

3.要知道，你在新闻标题中看到的数据常常是范围内的一个值。不论表述为误差范围（投票中的加/减）还是置信区间，研究发现通常只是一个范围内的估计值。如果一篇报纸上的文章说你最喜爱的候选人得票数为42%，那真实范围可能要加上或减去几个百分点。如果一份科学研究发现一种名字很好听的“益智”食品（包含鱼、浆果、绿色蔬菜）可能延缓阿尔茨海默病病情恶化的速度，置信区间能够提供其作用效果的范围。

4.哪怕得出了显著性差异的效应，也要**看一下效应量。**如果你在大洋里游泳，你可能会遭到鲨鱼攻击，但这种危险的概率很小。Discovery.com网站的一篇文章引用佛罗里达州大学国际鲨鱼袭击档案指出，实际上你“在海滩边溺水的概率是被鲨鱼袭击致死的概率的三倍”。你很容易因为最新发现有某种事物对你有害，或其他可能面对的危险而心神不宁——所以这就是为什么了解效应大小很重要的原因。

5.思考数据对你生活产生的影响。如果你生活在内布拉斯加州，从未计划出海，那你无须担心会被鲨鱼袭击，不论鲨鱼袭击人的概率有多大。（当然，除非你担心龙卷鲨）。如果你生活在非洲，你可能会更担心河马。根据盖茨基金会的说法，每年被河马袭击致死的人有数百人（与此相比，遭鲨鱼袭击致死的人数每年不超过12人）。仅仅因为某个研究发现具有统计影响——哪怕效应尺度很大，并不表示其对你每天生活会产生经济影响。

---

1. Michael Reilly, “Shark Attacks: What Are the Odds?” Discovery News website,August 2, 2010, http://news.discovery.com/animals/sharks/ shark- attacks- what- are- the- odds-20100802.htm.
2. Sharknado 3, Syfy website, accessed August 20, 2015, http://www.syfy.com/sharknado3.
3. Bill Gates, “The Deadliest Animal in the World,” April 25, 2014, Gatesnotes website, http://www.gatesnotes.com/Health/ Most- Lethal- Animal- Mosquito- Week.

第六章 非洲为什么会变小？歪曲与曲解

1544年，新晋制图师杰拉杜斯·麦卡托因为散布“异端邪说”被判入狱。显然，他的书信和游记引起了地方当局的猜忌，当局容忍不了他同情新教的言论。

他同时期有不少人都因为莫须有的罪名而被斩首、烧死或活埋，而麦卡托则在入狱7个月后获释，继续朝着当时欧洲顶尖制图师的方向前进。

你也许对麦卡托其人有所耳闻。要是没有，你应该见过他的作品——可能就挂在你小学教室黑板的旁边。尤其值得一提的是，他绘制了麦卡托世界地图（1569年版）——他不无夸耀地将其称为：“适用于航海的、较完善的新版地图。”（麦卡托说话总是直来直去。）

麦卡托的新版地图专为一件事情而绘制：帮助船员更为准确地绘制航线。若船只在某一航线上没有改变航向，那么该航道就用直线表示，这种新颖的方法使得船长在运送货物时更加轻松自如。

对于麦卡托来说，这种绘图技术需要一定的技巧，因为船员在立体地球表面的航行路线不能自动地转化成平面地图上的直线。不过麦卡托做到了，并因此名利双收。[注]

遗憾的是，新版地图在协助船员海上作业的同时，严重歪曲了全球陆地和国家的面积。船员为了获取直行航线要付出极大的代价，原因在于，麦卡托几乎歪曲了地图上所有陆地的面积。更糟糕的是，这种歪曲并不均衡：陆地距离赤道越远，歪曲程度越严重。



图6－1 麦卡托投影图。CC BY-SA 3.0协议文本授权许可。用户$200inaire创建于维基共享资源
（https：//commons.wikimedia.org/wiki/File:Mercator_Blank_Map_World.png#fi

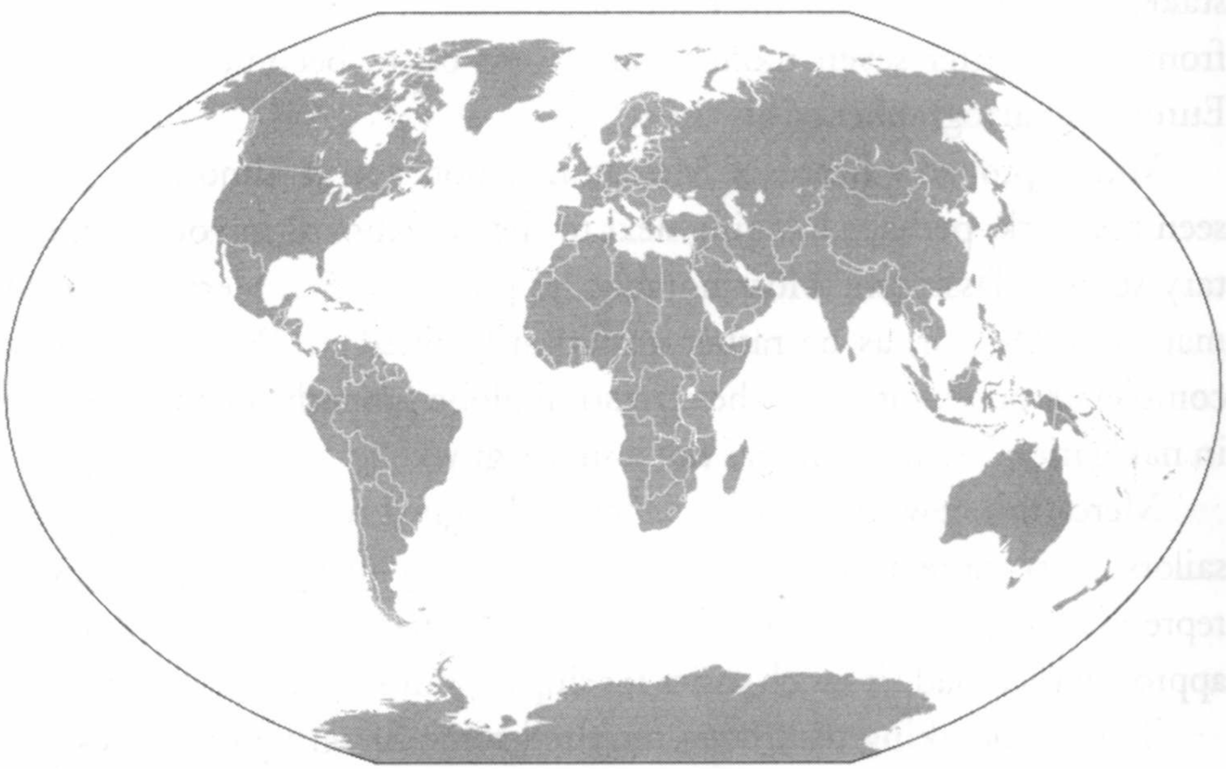


图6－2 为了对照，这是温克尔三重投影图（Winkel tripel projection）。CC BY-

SA 3.0协议文本授权许可。用户Hellerick创建于维基资源共享（https：//commons.wikimedia.org/wiki/File:1937_world_map_%28Winkel_tripel_projection%29.svg）

说得明白点，就是靠近两极的陆地会比靠近赤道的陆地看起来相对更大。这幅地图因“格陵兰难题”而普遍为人们（至少是为制图师们）所熟知。因为，在麦卡托地图上，格陵兰岛看起来似乎和非洲一样大。然而，事实并非如此，二者实际上相差甚远。

从面积上看，非洲面积几乎是格陵兰岛的14倍，然而非洲位于赤道（因此，歪曲程度较小），由于格陵兰岛大部分区域都在北极圈之内，所以，在麦卡托地图上，格陵兰岛的目测面积要比实际面积大得多。[注]

那么，这个错误会导致什么问题呢？除了造成几代学生困惑不解之外，我们所担心的是物体的大小（认知大小）会对现实世界的真实含义产生影响。麦卡托地图缩小了非洲大陆实际面积，有意或者无意地影响了人们从地缘、历史、政治等层面看待陆地的思维方式……想必你们也懂得这一点。不管你们是要计划度假、发动战争或者仅仅是为了争论欧洲与非洲孰轻孰重，陆地面积都举足轻重。

当然，这样的问题不是麦卡托地图所独有的——将三维数据转化到二维媒介上往往需要权衡再三——这就解释了为什么会存在60多种公认的地图投影。[注]但在我们身边无处不在的被歪曲数据之中，麦卡托地图是一个理想的范例。

读完本章你会发现，歪曲数据的情况几乎出现在生活的方方面面。不论你是在阅读年度报告，还是在点汉堡包，要是你略加了解数据是如何被歪曲和曲解的，那么你就可以做出更为明智的决定了。

---

1. Here’s a very basic explanation of how it works: picture a globe inside a cylinder. Now, project each point on the globe straight out until it makes a mark on the cylinder. Unroll the cylinder, and you have a Mercator map.

2. Another issue with the Mercator map is that, because it skews the size of objects so much at the poles, it's not uncommon for part of Antarctica to be cropped out. This has the effect of truncating the Southern Hemisphere, and therefore making the Northern Hemisphere appear larger in comparison.

3. And you're welcome to read all about them here: Melita Kennedy, "Understanding Map Projections," Environmental Systems Research Institute Inc., 2000,http://kartoweb.itc.nl/geometrics/map%20projections/understanding%20 map%20projections.pdf.

不要臆断

当然，要是你打算运用数据来评估人们的能力，你应该确保数据是真实的——完全真实，不是编造出来的，不是臆断出来的，甚至不是估算出来的。“维加拉诉加利福尼亚”（Vergara v.California）一案并没有做到这一点。案件中，法院认为加利福尼亚州的终身教职法使得贫穷及少数族裔学生受到不称职教师的拖累。[注]该案例围绕学生的测试成绩是否可以反映教师的授课效果及教师的授课效果如何影响其学生等假设展开的。

洛杉矶最高法院法官罗尔夫·特雷乌（Rolf Treu）在裁决书中写道：“加利福尼亚州请来为其辩护的专家伯利纳博士证实，加利福尼亚州1%~3%的教师极其不称职。”[注]

1%~3%这样的数据似乎有些言过其实了，那么这些数据是从哪里获取的呢？

加利福尼亚州专家证人戴维·伯利纳博士还引用了《Slate》杂志上一篇《读了才相信》的文章的内容，说道：“这是我的猜想，”[注]还补充说，“并没有相关的数据。”该数据仅是伯利纳基于访问“无数课堂”之后估算出来的。

伯利纳在《Slate》杂志文章的言论的基础上，他声称1%~3%是预估值。而且他似乎是一名受人尊敬的作家、美国国家教育研究院（National Academy of Education）会员，还身兼亚利桑那州立大学名誉教授。[注]

我们大力支持人人享有受教育权和平等权。（约翰担任Appleseed

董事会主席，Appleseed是一个倡导公平正义的国际非营利组织。）注 我们并不是在争论审判结果、由此带来的后果或者是伯利纳的资质。但是，我们认为此类的声明应该建立在确凿的数据之上，或者用其他的方式提出来。注

可能这样的事情曾经发生过，但不能将其作为一个默认的事实成为法官判决的依据。

如果不采用这种估算出来的数据的话，或许法官裁决的结果会有所不同。

再或者一个会影响27.5万名老师及600万名学生命运的案例会有截然不同的结果。

---

1. Full disclosure: Mike’s wife is a teacher. Yes, she has tenure. No, she’s not in California.
2. Beatriz Vergara v. State of California and California Teachers Association, Dept.58, No. BC484642 (August 27, 2014), http://studentsmatter.org/wp-content/uploads/2014/08/ SM_Final-Judgment_08.28.14.pdf.
3. Jordan Weissmann, “Fuzzy Math,” Slate website, accessed July 18, 2015, http://www.slate.com/articles/business/moneybox/2014/06/judge_strikes_down_california_s_teac
4. “David Berliner,” Arizona State website, accessed July 18, 2015, http://berliner.faculty.asu.edu/wordpress/. Oh, and in case you were wondering,the Sun Devil logo was designed by a former Walt Disney employee, and(according to some) supposedly resembles Mr. Disney in a not-so-flattering way. “Sun Devil Athletics,” the Sun Devils website, accessed July 18, 2015,http://www.thesundevils.com/ViewArticle.dbml?ATCLID=208256866&DB_OEM_ID=30300; Paul Lukas, “The Disney/ESPN Connection Suddenly Makes a Lot More Sense,” Uni Watch website, December 18, 2012, http://www. uni-watch.com/2012/12/18/ how- walt- disney- designed- the- umkc- mascot- character/.
5. See Appleseednetwork.org for details.
6. There were other issues with how Berliner’s testimony ended up as a key point in the judge’s decision. For example, Berliner claimed that he never used the words“grossly ineffective.” And his testimony (based on the Slate article) doesn’t seem to directly link bad teachers to low test scores.

柱状图和饼状图应该怎么画？

一幅恰当的图表能够讲述数据背后的故事，帮助你弄清楚数据之间的关系，从而让你做出更好的决定。若是图表不当，即使有无比翔实的数据集，也会让你觉得一团乱麻。

比如说，我们想弄明白每周运动时间与人们寿命之间的关系，那么我们会采用调研所得出的数据，记录参与者每周的运动量，及随着时间的推移对参与者死亡率造成的影响。[注]

每周运动时长	死亡率减少比率
0	0
0~7.5	20%
7.5~15	31%
15~22.5	37%
22.5~40	39%
40~75	39%
>75	31%

如数据所示，当每周运动时长从0增加到0～7.5小时这一区间时，

死亡率会降低20%。随着运动时长的增加，死亡率会持续降低，不过在此之后，运动的好处就开始趋于稳定。因为个人每周的运动时长如果在75小时以上，死亡率减少的比率就跌至每周运动时长在7.5～15小时对应的同等水平。

那么，我们把数据转化成图再看：

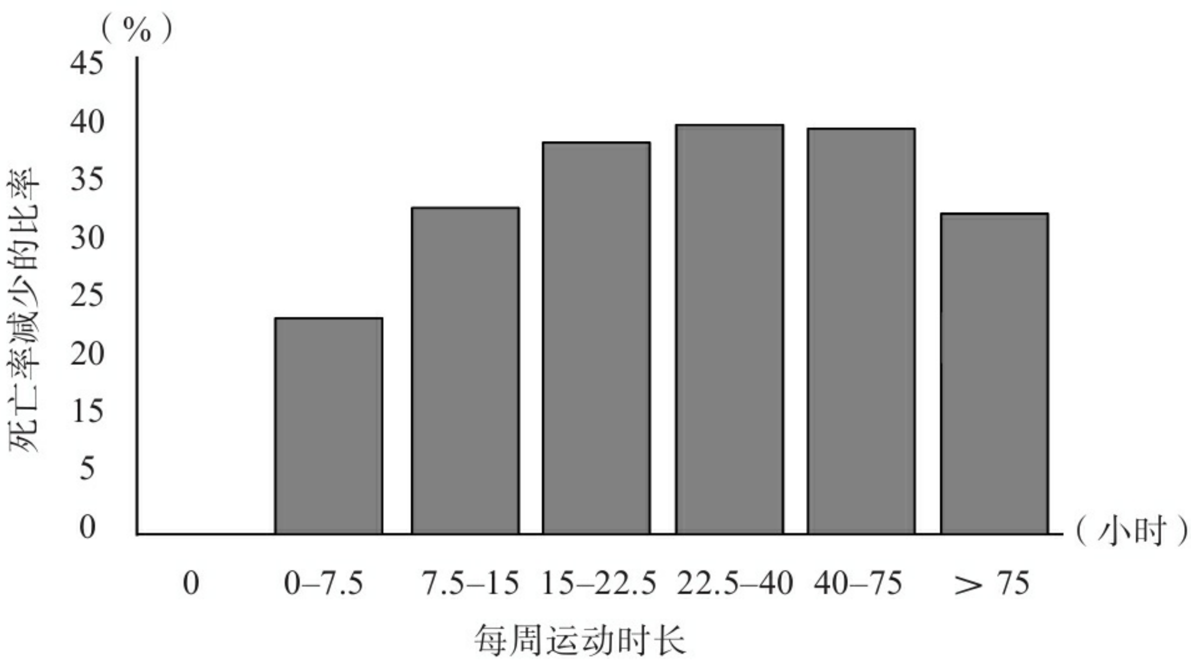


图6－3 运动时长与死亡率减少的关系

该图显示运动带来的好处是不断增加的，但随着运动量的进一步上升，运动带来的好处实际上不再增加了（后来还下降了）。

接着，来看看我们可以如何操纵数据。

假设你经营一个健身房，你想向你的临时顾客说明，他们只需稍加经常运动，就可能会延年益寿。你可以用这些数据，做如下图：

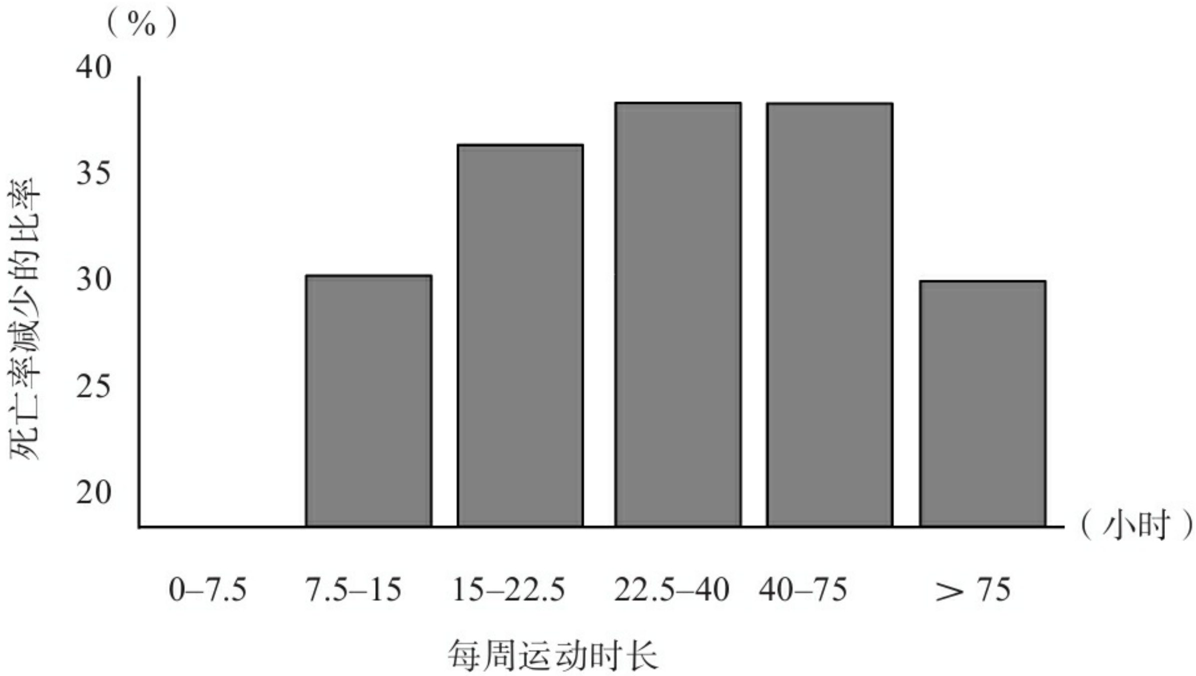


图6－4 我们操纵过的数据

这幅图跟第一幅图采用了完全相同的数据，我们所做的只不过是删去了0~7.5小时所对应的数据（随后所有图表都会这么做），将y轴（垂直数轴）的起点设为20（而不是0），在40处截止，稍稍高于数据的极值。看到其中的区别了吗？还有——或许更重要的是——看明白这幅图是如何使你想要每周在健身房花上7.5个小时以上的时间了吗？

如果我们确实想要强调运动的好处，可以如图6–5所示，使x轴（水平数轴）只显示到22.5～40小时处，以此消除持平及下降的趋势。

从这幅图来看，我们就制造了一种假象，似乎运动时间越长，好处越多。

现在，假设你的另一半一直央求你去健身房锻炼，增强体质、延长寿命。而你却更喜欢窝在沙发上看电视，因此你想使运动的好处最小化。你可以采用这些数据，绘制一幅图6–6这样的图：

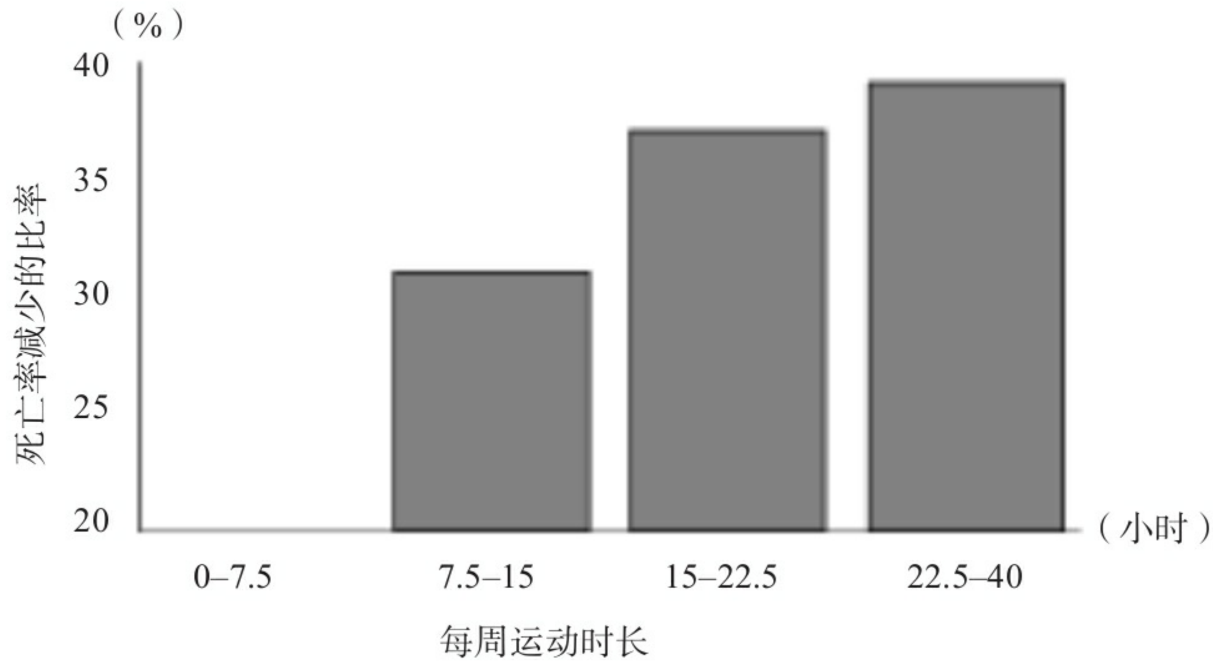


图6－5 x轴只显示到22.5~40小时处

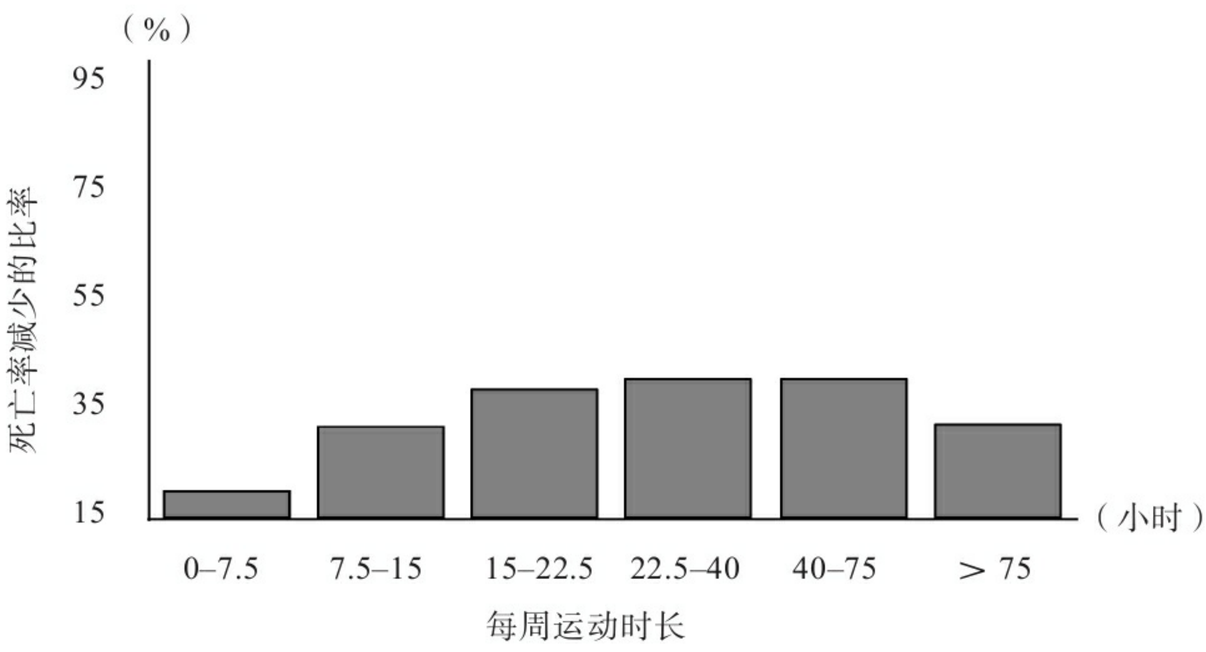


图6－6 降低了柱形条的高度的图

依然是完全相同的一组数据。在这种情况下，我们只是将y轴的起点和终点分别设为15和95，结果不仅降低了柱形条的高度，还缩小了它们之间人们可以感知的区别（降低的风险）。

当然，如果我们存心想迷惑你们——或者让你们误以为自己不应该运动，我们可以绘制图6–7：

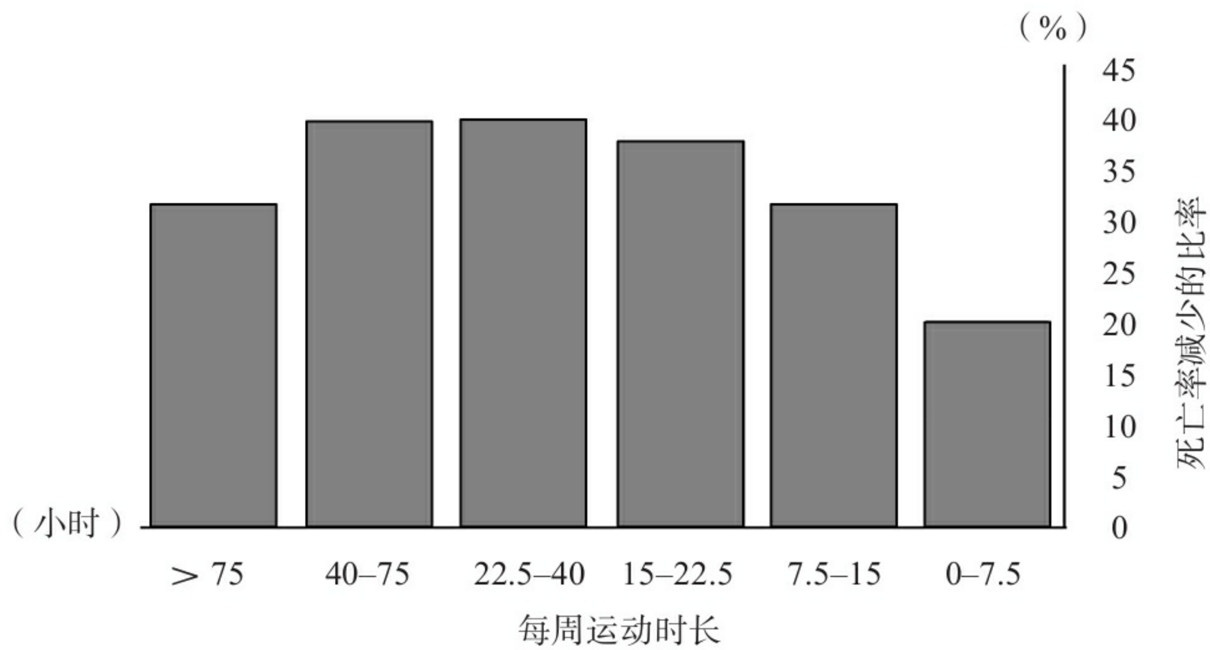


图6－7 x轴翻转到另外一侧的图

看明白我们做了什么吗？在柱形条图中，通常你看到的x轴是从左至右递增的。而一旦我们把x轴翻转到另一侧，除非你确实花时间去读懂所有的标签，不然你会觉得运动的好处似乎在缓慢增加之后减少。

在任何类型的数据图中，都需要留意这类“伎俩”。比如，饼状图通常用来表示不同小组在一个完整的数据集中所占的百分比，百分比总和为100%。所以，如果一幅饼状图中的数据之和为193%，真是会令人大跌眼镜的。注

正如x轴总是自左向右递增的，y轴也往往是从下往上递增的。但若我们翻转y轴，柱形条则向下延伸，（乍一看）呈负相关的关系。如图6–8所示：

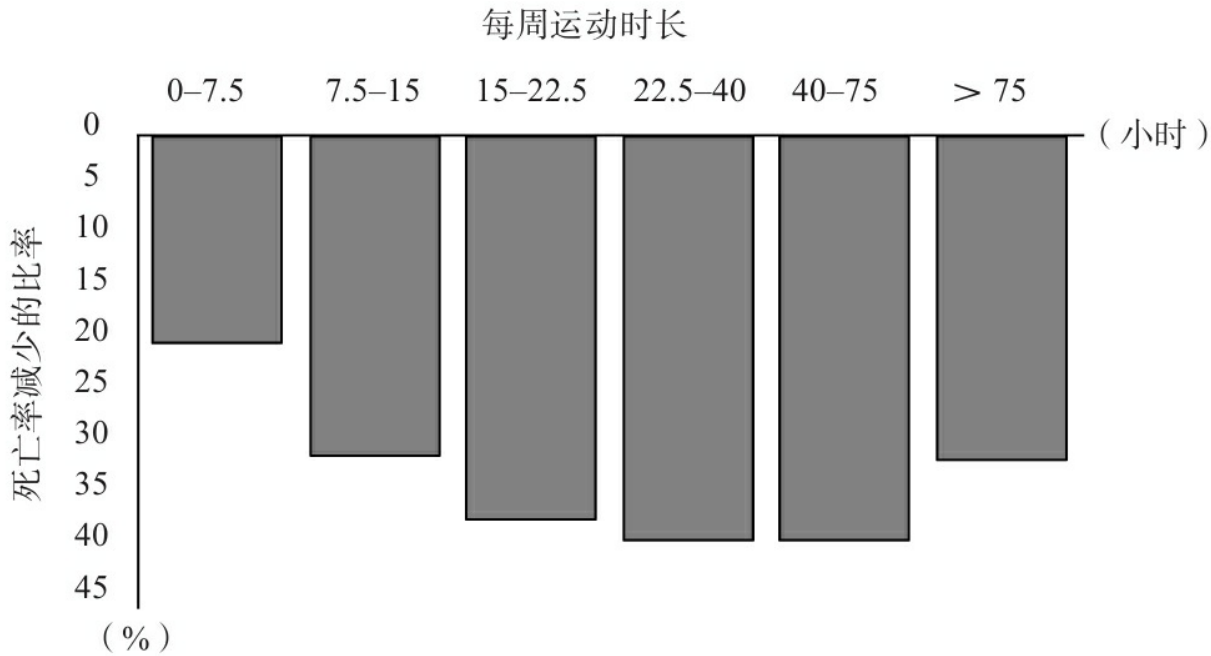


图6－8 翻转y轴的图

在图中，还需要发现的另一点就是，要看数据是累积式的还是增量式的。比如，倘使我们是从增量收益的角度来看这组数据会怎么样呢？每当运动量增加的时候，我们又会得到多少额外的好处呢？如图6–9。

区别很大，对吧？你只需花钱每周运动几个小时就能从中收获极大的好处，随后又开始下降——每周运动时长超过75个小时之后其实会给身体造成伤害。

另一方面，市场上最普遍的策略就是展示累积数据而不是增量数据，因为这能将负面消息转化成正面消息。举一个现实生活中的例子，一家知名科技公司通过展示手机的累积销量来突出其智能手机的销量。正如所料，柱形条从左至右持续上升，因为在累积式图表中每个季度的销量仅仅是和先前的总销量累积在一起的。而前一个季度的销量实际上下降了，但在图表上显示不出来。只要公司每个季度至少卖出几部手机，那么在累积式图表中，也看不出来销量下降，仅仅是增加得少而已。注

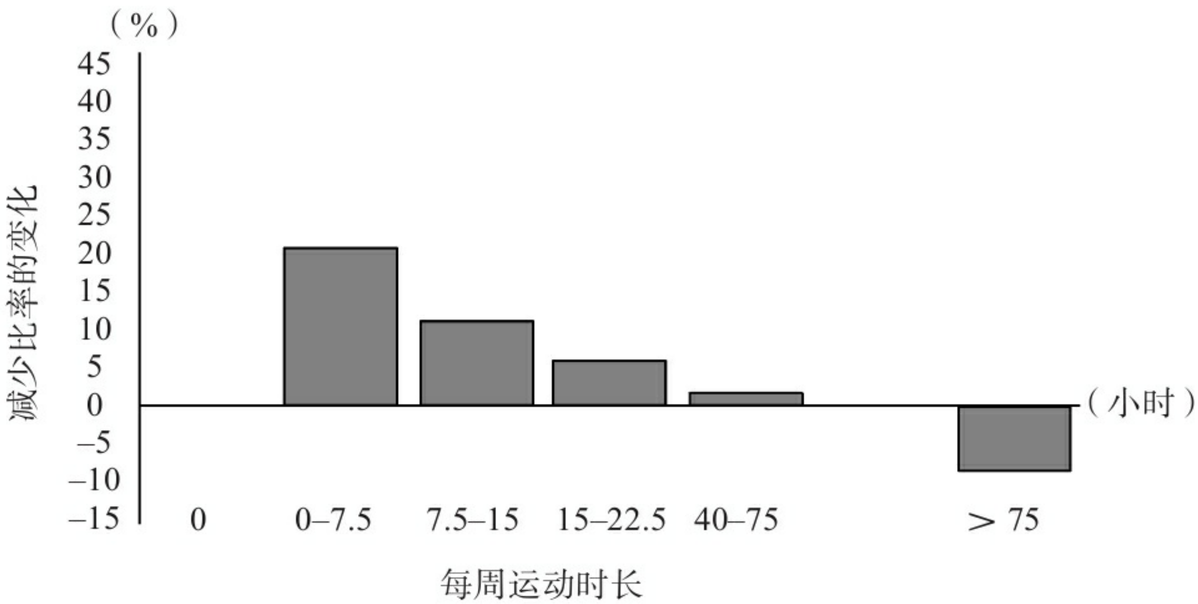


图6－9 从增量收益角度绘制的图

如果你有10分钟，并且你知道如何使用电子表格应用程序，你可以把完全相同的数据用多种不同的方式呈现出来。数据的呈现方式也会因数据要对读者所要传达的信息而存在巨大的差异。作为成熟的数据接收者，你们可以思考一下潜在数据是如何绘入图表之中的及它所要传达的真正含义。注

---

1. Hannah Arem, Steven C. Moore, Alpa Patel, Patricia Hartge, Amy Berrington de Gonzalez, Kala Visvanathan, Peter T. Campbell, Michal Freedman,Elisabete Welderpass, Hans Olov Adami, Martha S. Linet, L-Min Lee,and Charles E. Mathews, "Leisure Time Physical Activity and Mortality," JAMA Internal Medicine, published online April 6, 2015, doi:10.1001/jamainternmed.2015.0533.

2. Nathan Yau, "Fox News Makes the Best Pie Chart. Ever," Flowing Data website,accessed August 4, 2015. http://flowingdata.com/2009/11/26 fox- news- makesthe- best- pie- chart- ever/. The pie chart was aired on Fox Chicago, and the source was given as Opinion Dynamics.

3. David Yanofsky, "The Chart Tim Cook Doesn't Want You to See," Quartz website, September 10, 2013, http://qz.com/122921/ the- chart- tim- cook- doesnt- want- you-to-see/.

4. For a stunning look at how data can be captured, check out www. dear- data. com—"a year-long, analog data drawing project" by two extremely talented information designers: Giorgia

Lupi and Stefanie Posavec, “Dear Data,”accessed June 7, 2015, http://www. dear- data.com/.

微妙的圆

在柱状图中，你往往只需比较柱形条的高度（或者水平图表的宽度）。但如果你手头的图表使用的是圆，那事情可能会变得有些微妙。

以美国劳动统计局（Bureau of Labour Statistics）的数据为例，我们实际上可以逐时细分普通美国民众是如何度日的。(注)来看一下我们每天花在看电视与社交活动的时间对比——毫不意外，结果大幅偏向看电视。2014年，我们平均每天看电视的时间为2.82小时，而用于社交活动的时间为0.71小时。

那么，你画两个圆，用每个圆的面积代表时间总量。得到下列对比图：

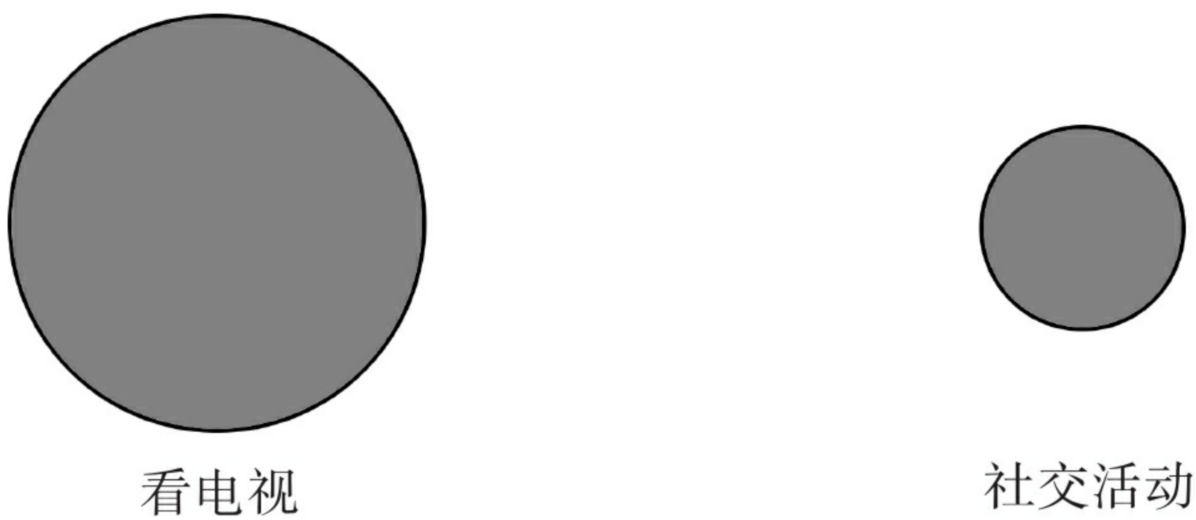


图6－10 用面积表示时间总量

所以我们看电视的时间更多，不过也没有多太多，是吧？

现在，依照同一组数据，我们再画两个圆——这次用圆的直径来表

示时间总量。新的图如下：

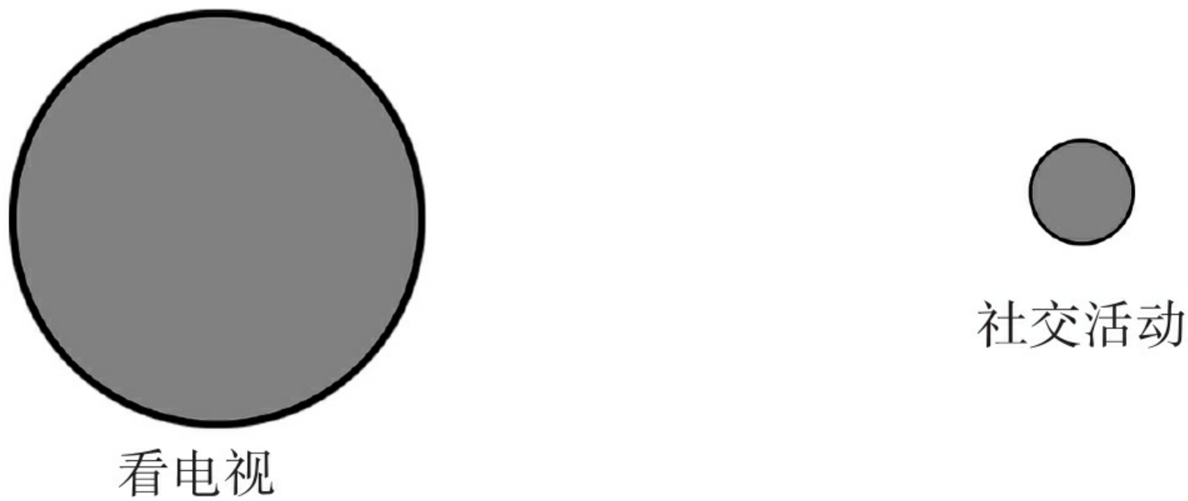


图6－11 用直径表示时间总量

从这幅图来看，似乎我们太多的时间都在电视机前度过了。

这是为什么呢？因为数学在作怪。简言之，一个圆的直径和面积并不是一回事。任何一个中学生都能告诉你，圆的面积等于π乘以半径的平方（$S=\pi r^2$）。因此，当圆的直径变为原来的2倍时，圆的面积实际上为原来的4倍。

正如在这个例子中看到的，用圆的直径表示数据比圆的面积来表示数据更能为你们呈现出其中的差异。从统计学层面上来讲，两种方法本身都没有错。有人认为用面积表示数据更加清楚明了，因为用直径来表示的话往往会令人一头雾水。毕竟，你既然能采用直径（一条直线）表示，为什么不用柱状图呢？

无数的时间（及诉讼费）都用在对这些现实问题的争论上了。我们的目的不是提供一个定论，仅仅是为了让你们下次看到用圆表示的数据时，停止争论，多问问为什么。注

1. “American Time Use,” Bureau of Labor Statistics website, http://data.bls.gov/ cgi-bin/surveymost?tu.
2. For a discussion of what type of visualization to use, check out our blog post on the topic at http://www.johnhjohnsonphd.com/blog2/blog/2015/4/18/a-guide-to-data-visualization.

图表让人看起来更值得信任

视觉图形还可以让数据看起来更加可信。正如《哈佛商业评论》所提到的，引用了康奈尔大学的研究成果，“当有关新药疗效的声明以文本形式呈现出来时，67%的研究对象称他们会相信，但是当同样的声明以文本和一幅简单的图表一起呈现时，则有97%的人会相信”。[注]这反映出我们在某些案例中所遇到的偏见，在这些案例中，效果不是用数据衡量，而是通过视觉或者其他方式表现出来。（这类似于“教条主义”。在拉丁语中，ipse dixit意为“他亲口所述”，这个词通常会用在有一位专家说某件事是真实的，而这件事情之所以真实是因为专家这么说了。就好比你过去常常问父母为什么你必须上床睡觉，他们回答说“我们说了算”一样。实际上，将来你们的孩子问你们为什么非得做某件事，我们极力推荐只用“ipse dixit”回应。）

下次你看电视的时候，记录一下有多少商业广告中会有身着白大褂的人出现，这是医学或者科学专家典型的视觉符号。暗含的意思一清二楚——这是你们应该信任的人。

我们赞成，往往好的专家比普通人更加专业，更加值得信赖，而且可以凭借他们的学识帮助我们加深对重要议题的理解。然而仅仅披上一身白大褂（或者手术服，抑或一套正装）并不能让人摇身一变成为专家。

记住，每当一幅图像出现在你眼前，那都是有人有意为之，以达成某一目的。比如，餐馆菜单展示出来的图标和照片能够带动销售额上涨多达30%。[注]

然而我们可能需要仔细观察一下这些视觉图形。正如一位数据记者

在为《卫报》的撰文中精妙绝伦的阐释，“将图表及数据可视化，在绝大多数情况下仅被用作呈现最终结果的媒介。这么做的结果便是人们边读文本边想‘我不赞成’要比边看精心准备的示意图或线形图边这么想要来得更加理所当然”。(注)

前文已经提过，这里再次重申：作为一名受过训练的数据接收者，你们的目标应该是在获取真实的信息之前，不断深入挖掘。不过这么做往往并不容易。这是为什么呢？知名广告人兼作家鲍勃·霍夫曼（Bob Hoffman）在给我们的一封电子邮件中指出：“在商界中见到的大多数研究报道不会把主要数据、问卷途径及调查方法透露给我们。相反，他们给我们一幅图表或者若干个结论。结果，我们很难知道研究是否是在恰当的方式下开展的。”

---

1. “To Make Your Claim More Believable, Simply Add a Graph,” Harvard Business Review, https://hbr.org/2014/10/to- make- your- claim- more- believable- simply- add-a-graph. The original research is here: Aner Tal and Brian Wansink,“Blinded with Science: Trivial Graphs and Formulas Increase Ad Persuasiveness and Belief in Product Efficacy,”Public Understanding of Science, October 15,2014, doi: 10.1177/0963662514549688.
2. “Menu Engineering: How to Raise Restaurant Profits 15% or More,” Menu Cover Depot website, accessed June 16, 2015, http://www.menucoverdepot.com/ resource- center/articles/ restaurant- menu- engineering/.
3. John Burn- Murdoch, “Why You Should Never Trust Data Visualisation,”Guardian website, July 24, 2013, http://www.theguardian.com/news/datablog/2013/jul/24/ why- you- should- never- trust-a- data- visualisation.

对所有数据一视同仁导致曲解数据

另一种曲解数据的手段是哪怕数据并不对等，依然对其一视同仁。

假如你有一个反对拿动物做实验的朋友，他发给你一份名单，上面列出了拿动物给其产品做实验的公司，并要求你抵制这些公司。[注]

如果纸上就列出了一排公司名，或印满了公司商标，那么就意味着这些公司要因为拿动物做实验承担相同的责任。但是事实果真如此吗？

所有公司拿动物做实验的次数是一样的吗？它们都使用了同种类型的化学试剂吗？它们的实验对象是同种类的动物吗？

实际上，这些公司在用动物为自己的产品做实验时采取的方式不可能完全对等，但是一旦它们被简单地罗列在一张纸上，就会被误认为它们是一样的。就像给你的妻子发信息说你刚吃了点“炸薯条和胡萝卜”。你每样都吃了10个呢？还是吃了2根胡萝卜、50根炸薯条？对所有的数据一视同仁，也是曲解数据的一种形式。

在主持一场有关气候变化的辩论时，电视脱口秀主持人约翰·奥利弗（John Oliver）极为巧妙地阐释了这一观念。[注]在典型的电视辩论会中，话题的正反方各有一个或者两个辩手，而奥利弗则不同以往，决定正反方辩手的人数要体现出持相应观点的人数。因此，他邀请了97个人作为正方，认为气候变化是真实的——3个人为反方，持反对观点。仅凭借这一简单的举动，他彻底改变了普通人看待这场辩论的视角——我们觉得，这或许正中他的下怀吧。

---

1. “PETA’s Beauty Without Bunnies Program, Companies That Do Test on Animals,” PETA, accessed June 7, 2015, http://www.mediapeta.com/peta/PDF/companiesdotest.pdf.

2. Last Week Tonight with John Oliver, Episode 3: “Climate Change Debate,” Last Week Tonight video, 4:27, HBO, May 11, 2014, http://www.hbo.com/ last- weektonight- with-john-oliver/episodes/01/03-may-11-2014/video/ climate- changedebate.html?autoplay=true.

体会有和仅有的区别

仔细玩味以下句子：

22%的鲨鱼攻击事件是致命的。

仅有22%的鲨鱼攻击事件是致命的。

加上“仅有”（only）一词，我们就不会很重视其后的数据——这是歪曲数据的另一手段。正如劳动局的统计员所阐释的，“仅有”一词激发了一种对某些与众不同的事物的不切实际的期待。[注]

增删文字也是常见的曲解数据的手法。

鲍勃·霍夫曼曾在其博客Ad Contrarian里反复提到一个数据，60%的人说他们使用QR码[注]。[注]霍夫曼说：“这一统计数据明显是胡说八道，然而刻板的人们却似乎把它当真了。任何生活在现实生活中的人都明白，没有人使用QR码。”[注]

那么60%源自哪里呢？或许正如霍夫曼的推断，它可能是曾经使用过QR码的人所占的百分比。在没有上下文对数据进行界定的情况下，霍夫曼指出：“以一种取巧的方式陈述真相，事实由此被彻底歪曲了。”

启示——如果你只看表面数据，你可能把握不了全局。除非开口去问，否则你不知道数据是否被歪曲或者被删减了。

---

1. Bryan Beverly, “3 Old Tricks for the Analytics Hall of Shame,” All Analytics website, November 25, 2013, http://www.allanalytics.com/author.asp?section_id=1828&doc_id=269454&f_src=allanalytics_sitedefault&utm_source=dlvr.it&utm_me

2. QR码是二维码的一种，其内容可以被快速读取。——编者注

3. A quick response (QR) code is a type of bar code that users can scan using their smartphones to get video and other content.
4. Bob Hoffman, “How Marketers Lie to Themselves,” Ad Contrarian blog, April 20,2015, http://adcontrarian.blogspot.com/2015/04/ how- marketers - lie-to-themselves.html.

油表显示油箱为空为什么还能开个几里路？

有时，故意歪曲数据不是为了误导你，而是为了帮你。

如果你驾驶过一辆燃油不足的汽车，或许会注意到一个奇怪的现象——即便在油表指针显示油箱为空的情况下，车辆通常还可以继续行驶。从这个数据（油表）来说，你应该是没有多余的汽油了，但实际上却还有。

假设油表坏了，这还情有可原——汽车制造商知道，大多数人不喜欢燃油耗尽的情况。据一则美国广播公司新闻报道，福特（Ford）、通用（GM）和克莱斯勒（Chrysler）均为美国国内司机提供了一个“缓冲器”（尽管对他国司机则不完全尽然，他们可能希望油表读数更为精确）。[注]

那么汽车制造商做了什么呢？它们这么呈现数据：让你汽车的燃油量看起来比实际燃油量少。所以，下次你的油表指针指到E——并且汽油没有耗尽——记得要感谢造你这辆车的人。

---

1. John Stossel, “Running on Empty,” ABC News website, June 5, 2008, http://abcnews.go.com/2020/Stossel/story?id=3989000.

造假与失误

你听过17000名英国男子怀孕的消息吗？

在一封给英国医学杂志（British Medical Journal）的信中，三名医生引用数据表示17000多名男士通过英国的全民医保（National Health Service）接受了住院产科服务。㊟

显然，这是假的。男人们连换尿布都不愿意，更别说让他们怀胎十月了。男人怀孕最有可能是因为医疗编码错误。换句话说，有人在医生诊断的时候输错了数据。

下面是一些更为典型的例子：

•鱼的那些事儿——从商店或者餐馆购得的金枪鱼样品，其中59%都被贴错了标签；常常，金枪鱼实际上是玉梭鱼，属蛇鲭鱼科，可能会引发一些身体状况，说得正式一点，即肠胃不适。㊟

•就为了评（假）星级——在一次秘密行动中，纽约州总检察官办公室人员发现很多公司为了生意在Yelp（美国最大点评网）、Google Local（谷歌本地）等网站上编写虚假网评——该做法名为“草根营销”（astroturfing）。

•名不副实的零脂肪——美国食品及药品管理局（FDA）批准，每份食物脂肪量低于半克仍可以称作“零脂肪”。因此，如果你每天吃多份“零脂肪”食物，你就很轻易地就摄入了几克脂肪。㊟

•棘手的单元格——正如彭博商业资讯（Bloomberg Business）所称，它是“改变历史的Excel表格错误”。㊟两位哈佛大学的经济学家

——卡门·莱茵哈特（Carmen Reinhart）和肯尼斯·罗格夫（Kenneth Rogoff）——因为错误的推理上了头条，当时，在一篇调查政府债务对经济增长的影响的论文中，他们绘错了电子表格，忘记在其中一组运算中插入5个空行，结果使得关键结果由+0.2%变成了–0.1%。（《经济学人》还指出了其他造成运算大错特错的错误。）

---

1. Lauren Brennan, Mando Watson, Robert Klaber, Tagore Charles, “The Importance of Knowing Context of Hospital Episode Statistics When Reconfiguring the NHS,” BMJ 2012; 344:e2432.
2. Kimberly Warner, Walker Timme, Beth Lowell, and Michael Hirshfield,“Oceana Study Reveals Seafood Fraud Nationwide,” Oceana website, February 2013 http://oceana.org/sites/default/files/reports/National_Seafood_Fraud_Testing Results_FINAL.pdf.
3. “Guidance for Industry: A Food Labeling Guide (9. Appendix A: Definitions of Nutrent Content Claims),” U.S. Food and Drug Administration website, January 2013,http://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/
4. Peter Coy, “FAQ: Reinhart, Rogoff, and the Excel Error That Changed History,” Bloomberg Business website, April 18, 2013, http://www.bloomberg.com/bw/articles/2013-04-18/ faq- reinhart- rogoff- and- the- excel- error- that- changed- history.

错误数据抹掉股市1360亿美元

还有错误数据抹掉股市1360亿美元的时候。

2013年4月23日，美联社推特账号发布了以下消息：

“重磅消息：白宫发生两起爆炸事件，巴拉克·奥巴马负伤。”

投资者惊慌失措，股票价格骤然下跌，标准普尔500指数（S&P500）仅在两分钟内就损失了1360多亿美元。

一切都是因为一个恶作剧而起。

当人们意识到推特消息不实时，股票市场迅速恢复。没有爆炸事件发生，奥巴马总统也没有受伤。但是如果你持有股票或者共有基金，在这惊心动魄的几分钟里，虚假数据会对你的证券投资产生极为真切的反响。

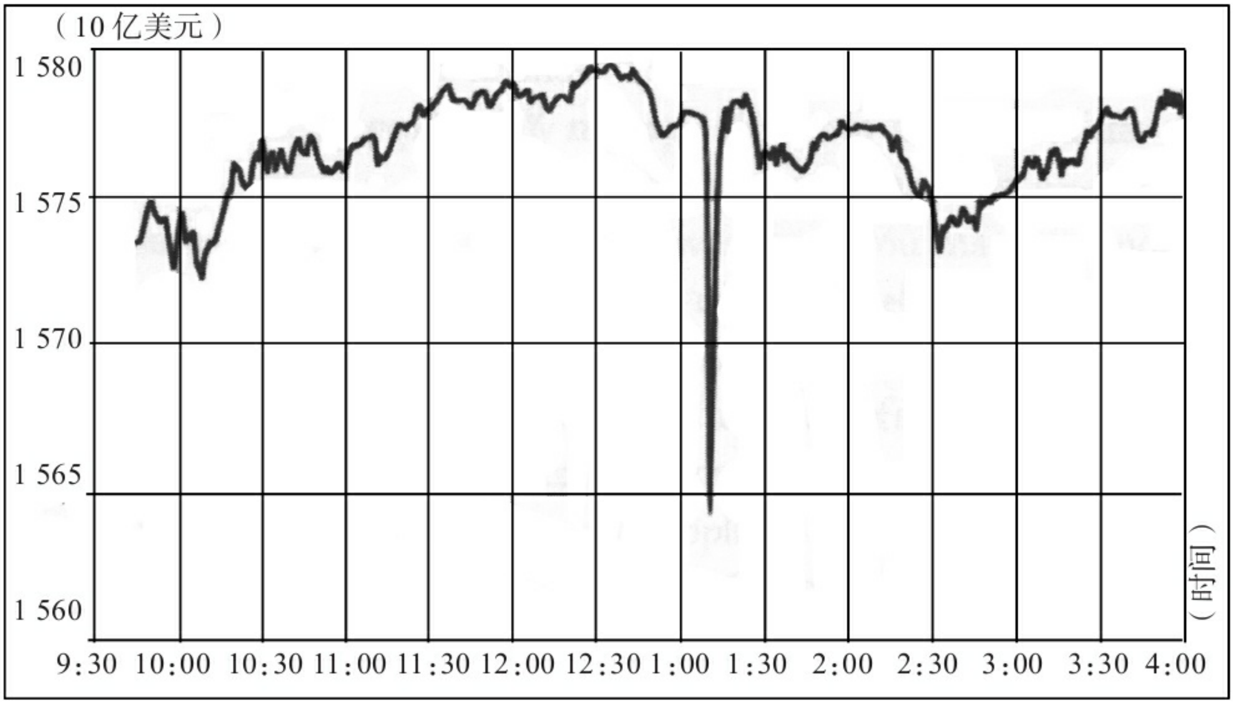


图6－12 可以看出下午1点之后，陡降的标准普尔500指数（S&P500）

不要相信维基百科

最后，我们来谈一谈广受好评、历史悠久的维基百科（Wikipedia），以此作为歪曲数据的收尾。很多大学生的学期论文都是在该网站的帮助下完成的，用维基百科自己的话来说，它们提供“公开的、可编辑的内容”。[注]

作为一个聪明的数据接收者，看到最后这句——“公开的、可编辑的内容”，应该会顿一下。如果任何网民都可以在这个全世界最受欢迎的网站更新内容，究竟凭什么要相信上面的信息？答案很明了：不能相信。[注]正如该网站所注明的那样，“尽管网站上有些内容具有很高的学术价值，但也存在被公认为彻头彻尾的垃圾信息”。[注]

所以，如果你打算依赖维基百科，要确保核实过上面内容的来源。

---

1. “Wikipedia: About,” Wikipedia, accessed June 13, 2015, https://en.wikipedia.org/wiki/Wikipedia:About.
2. That said, even when the data does come from a trusted media source, you should at least recognize the fact that a few large media conglomerates are responsible for much of the news and entertainment you consume on a daily basis, which could have a significant effect on what you see and hear.
3. “Ten Things You May Not Know About Wikipedia,” Wikipedia, accessed June 13, 2015, https://en.wikipedia.org/wiki/Wikipedia:Ten_things_you_may_not_know_about_Wikipedia#You

过了保质期的食品能不能吃？

在这一章，到目前为止，我们的关注点都集中在了数据被歪曲的例子上——个人、公司或组织向你提供了不太真实的数据。

另一方面，你需要确保自己没有曲解真实的数据。

比方说，食品一过包装袋上的保质期，有的人就会立刻把它们丢掉，如果你是其中一员，也许每年你会浪费掉数百甚至更多美元。那是因为大多数人所理解的食品保质期根本不是保质期。

据美国农业部（USDA）下属单位食品安全监察局的说法，许多食物过了包装上的日期之后依然可以食用。实际上，日期代表的不是有效期，而在大多数案例中，代表的则是产品质量峰值的日期。有些州规定某些食品需要注明日期，但是一般情况下，联邦政府规章不要求注明日期，婴儿配方奶粉除外。

下面是一些直接摘自美国农业部的网站的参考资料：[注]

•销售日期（Sell-By date）是要告诉店家展销商品需要多长时间。你们应该在过期前购买商品。

•在此日期或此日期前食用最佳（Best if Used By or Before），该日期用来给那些想要获得最佳口感及质地的消费者提供建议。它不是购买日期或者安全食用日期。

•保鲜期（Use-by date）即产品质量能够保持峰值的最后日期，该日期由产品制造方确定。这就是说，美国农业部通常会建议人们根据保鲜期食用食品，也建议将鸡蛋置于冰箱的冷冻室里，而不是放进冰箱就

行。

因为这三种类型日期——三者都不是真正的保质期——人们很容易对隐藏在数据背后的含义困惑不解。实际上，根据美国自然资源保护委员会（NRDC）引自食品营销协会的研究结果来看，对日期的困惑“致使90%的美国人扔掉了本不该扔的食品”。(注)

所谓“保质期”这一说法在小数据领域为人们上了精彩的一课，因为它阐明了解数据所代表的含义有多重要。食品制造商没有歪曲“保质期”，但是人们理解数据的方式则可能使很大一部分预算付诸东流。(注)

---

1. “Food Product Dating,” United States Department of Agriculture website,accessed June 16, 2015, http://www.fsis.usda.gov/wps/portal/fsis/topics/ food- safety- education/ get- answers/ food- safety- fact- sheets/ food- labeling/ food- product- dating/ food- product- dating.
2. “Environmental Issues: Food and Agriculture, the Dating Game,” Natural Resources Defense Council, accessed June 16, 2015, http://www.nrdc.org/food/expiration- dates.asp.
3. Read our full blog post on this topic— including some details on how much money wasted food might cost us—“Everydata of Expiration Dates,” the John H. Johnson, PhD Blog, May 18, 2015, http://www.johnhjohnsonphd.com/blog2/blog/2015/4/18/expiring-soonbut-so-what.

确凿无误的消息也可能被误读

当谷歌宣布收购一家名为Nest的恒温器公司时，有人认为他们可以通过购入交易名为NEST的公司的股票小赚一笔。仅在一天之内（2014年1月14日），NEST的股票价格就上涨了1900%。

对于企图一夜暴富的人来说，很遗憾，NEST不是Nest公司的股票名，而是Nestor公司的股票名称，而Nestor是一家销售交通执法系统的公司。恒温器公司Nest的股票并没有公开交易，尽管自2015年起，它归谷歌建立的控股公司Alphabet所有。2009年，Nestor进入破产管理流程，没有资产剩余。

这一数据是精准的。谷歌收购Nest的消息是确凿无误的，然而投资者没有去核实，结果购买了仙股[注]而不是谷歌新公司的股票。Nestor股票价格确实下降了——尽管降速没有涨速快。甚至在2014年1月14日这一天即将结束时，交易额比当天开盘时高大约400个百分点。

---

1. 仙股最初源于香港的股市，是特指市值跌至1元以下的股票，在英语中被称为penny stock。——译者注

1/4磅牛肉汉堡与1/3磅牛肉汉堡哪个大？

美国人不擅长数学，或者说，数学非常烂。在一项研究报告中，美国人的数学水平在23个国家中排名第21。[注]这或许就能解释为什么A&W餐厅的汉堡会失败了。

据《纽约时报杂志》报道称，在20世纪80年代，A&W连锁餐厅欲与麦当劳及其著名的“1/4磅牛肉汉堡”争个高下。[注]因此，A&W决定推出1/3磅牛肉汉堡。顾客认为后者口味更佳，结果却销量惨淡。显然人们以为1/4磅的汉堡比1/3磅的汉堡大。

为什么他们会觉得1/4大于1/3呢？因为4大于3。

是的，这是真事，我没开玩笑。

人们误解汉堡的大小，是因为他们看不懂分数。

在研究中，我们发现很多人在比较数字、分数及百分数时很吃力。一项著名的研究发现，在“1万人中有1286个人因癌症去世”与“100个人中有24.14个人因癌症去世”这两种情况当中，人们会认为前者的癌症的风险更高一些。[注]（数据基于以往研究，研究时还要求天真的参与者估算死亡率。）

如果你自己算一下的话，轻易就能看出1286/10000小于24.14/100。问题在于许多人不知道如何计算，或者他们只是简单地猜想首个选项必定风险更高，因为1286大于24.14。

你该如何解决这个问题呢？除了时光倒流，当一名三年级的数学老师之外，你还可以在想让人们比较数据大小的时候，先把数据格式统一

一下。

我们明白，一眼看出24.14/100大于1286/10000对很多人来说并非易事，但是也难以辩驳2414大于1286的事实。

---

1. The study included a “nationally representative sample of 5,000 adults between the ages of 16 and 65,” with “similar nationally representative samples” from other countries. On the numeracy test, “Compared with the U.S. average score,average scores in 18 countries were higher, in 2 countries they were lower, and in 2 countries they were not significantly different.” So, while the U.S. was listed 21st out of 23, numbers 19 (Ireland) and 20 (France) were not “significantly”higher: Madeline Goodman, Robert Finnegan, Leyla Mohadjer, Tom Krenzke,and Jacquie Hogan, “Literacy, Numeracy, and Problem Solving in Technology Rich Environments Among U.S. Adults: Results from the Program for International Assessment of Adult Competencies 2012: First Look,” U.S. Department of Education, National Center for Education Statistics, NCES 2014- 008, October 2013, http://nces.ed.gov/pubs2014/2014008.pdf.
2. Elizabeth Green, “Why Do Americans Stink at Math?” New York Times Magazine website, July 23, 2014, http://www.nytimes.com/2014/07/27/magazine/why-do- americans- stink-at-math.html.
3. Kimihiko Yamagishi, “When a 12.86% Mortality Is More Dangerous Than 24.14%: Implications for Risk Communication,” Applied Cognitive Psychology 11 (1997): 495– 506.

如何明智地接收被歪曲（或可能被歪曲）的数据

数据被歪曲或者曲解的方式不一而足，但是想要做一名明智的数据接收者，你可以从下面5件事做起。

1.碰到图表时，仔细观察x轴和y轴。简单调整比例、高度或单个（或者两个）数轴的其他方面来传达出完全不同的信息，这对有些人来说轻而易举。比例的起点终点在哪里？数字在增长还是在下跌？图表是否涵盖了所有相关数据？这些都是你可以问的问题。

2.留意语言表达。文字究竟表达了什么？若一家工厂称“最近”都没有事故发生，这意味着什么？由谁来界定什么才算“事故”？你用不着逢人就问，但是弄明白人们的言外之意常常和听人们口中的话同样重要。

3.核实信息源。仅仅因为你在网上看过（从老板那里听过，或者在报纸上读到过，再或者看过电视报道……）并不意味着它真实可靠。地球虽然是圆的，但假如你相信地平说协会的说法，你就会对此不以为然。[注]

4.确保数据无误。正如《电讯报》（Telegraph）摘引咨询公司F1F9的报告所提到的，“几乎有1/5的大企业因为电子表格出错而蒙受财产损失”[注]。工作做完要仔细检查一遍，检查完毕后，再检查一遍。

5.正确解读数据。你买入的股票，真的是你想要买入的那只吗？你知道鸡蛋盒上保质期的真正含义吗？你会对分数、小数点、百分比之间的区别感到困惑吗？有时数据没有错，但是会因为人们的草率、无知等因素被误解。

---

1. Actually, we found two flat earth societies in a quick search, both of which are called the

Flat Earth Society: http://www.theflatearthsociety.org/cms/; http:/www.tfes.org/. Both accessed September 1, 2015.

2. Rebecca Burn- Callander, “Stupid Errors in Spreadsheets Could Lead to Britain’s Next Corporate Disaster,” Telegraph website, April 7, 2015, http://www.telegraph.co.uk/finance/newsbysector/banksandfinance/11518242/ Stupid- errors-in-spreadsheets- could- lead-to- Britains- next- corporate- disaster.html.

第七章 筛选数据需要“摘樱桃法则”：疯狂筛选

假设回到1996年，你有一个活蹦乱跳的小女儿。你刚把她哄睡着，喘口气坐下，想好好看几分钟电视。

电视中出现了一则商业广告，画面中有罐装婴儿食品，还有新鲜的苹果和胡萝卜的照片。父母总想把最好的留给下一代，所以你怀着一颗慈爱之心留意着广告的内容。㊟

你看到了乐呵呵的婴儿，看到了罐装婴儿食品，听到了播音员说这种婴儿食品如何如何好，还说，“推荐婴儿食品的儿科医生中，有4/5推荐了嘉宝（Gerber），想知道为什么吗，请随时来电咨询，24小时在线”。㊟

嘉宝在给消费者的电话留言中反复强调这一论点。当人们拨打此公司的电话时，会听到电话那头的录音：“推荐婴儿食品的儿科医生中，有4/5推荐了嘉宝。”

有4/5的儿科医生推荐。听起来似乎这款产品获得的支持率很高，难道不是吗？

上面这句广告词对那些被孩子搞得睡眠不足的家长尤为有效。如果4/5的儿科医生都推荐了嘉宝，那就意味着仅有1/5的医生未推荐，是吗？并非完全如此。

事实上，该公司调查的所有儿科医生中，仅有12%的人推荐嘉宝。那么“4/5”这个说法从何而来？

问题在于嘉宝对数据进行了筛选。你可能听过“摘樱桃法则”。实际上，你可能自己也这么做过。摘樱桃法则意味着你从数据中挑选吸引人眼球的例子来证明自己的观点，同时无视那些可能跟自己观点相悖的数据。嘉宝绝对不是唯一一家通过筛选数据来营销的公司。但是在这个特别的例子中，联邦贸易委员会（FTC）——美国政府机构，职责包括防止虚假、不公正的商业活动——深入调查了嘉宝广告中的论点，并将其定性为“虚假宣传且误导大众”。[注]

嘉宝并没有说4/5的儿科医生推荐了他们品牌的婴儿食品。而是说，推荐婴儿食品的儿科医生中，有4/5推荐了嘉宝。公司只采纳了推荐婴儿食品的儿科医生的回答，从而从特意拣选出来的数据样本中选择了对自己最有利的那部分。

进行这样的区分十分重要，因为尤其在当时，很多儿科医生因为顾虑到婴儿食品含有糖分、添加剂及其他成分的问题，并没有对其进行推荐。在这个例子中，参与嘉宝问卷调查的儿科医生中，超过1/4根本没有推荐婴儿食品。[注]若嘉宝将这部分儿科医生算入其调查结果中，“4/5”这样的说法就站不住脚了。

值得赞扬的是，嘉宝很坦率地承认自己是如何利用儿科医生的样本并筛选数据的。根据我们研究过的案例，嘉宝通过告诉消费者它们只谈论推荐婴儿食品的儿科医生的说法，以此证明自己的观点是正确的。

但是嘉宝在其他方面也对数据进行筛选，而且使用的方式较为隐蔽。

确切地说，嘉宝一开始对562名儿科医生进行了问卷调查。据联邦贸易委员会称，“其中的408名反映说他们每周至少会向自己的患者推荐一次婴儿食品”。[注]基于“推荐婴儿食品的儿科医生中，有4/5都推荐了嘉宝”这一说法，你会认为这408名儿科医生中，大多数人都认可了嘉宝。

事实可不是这样。看到了吧，嘉宝不仅挑选最优数据，它还优中选优。

因为在推荐婴儿食品的408名儿科医生中，有332名没有推荐某一特定品牌。他们只说，患儿每周至少食用一次婴儿食品。

换句话说，嘉宝没有将没有推荐某一特定品牌婴儿食品的儿科医生算在内，而只将推荐婴儿食品的医生（A）及推荐某一特定品牌的医生（B）计算在内。

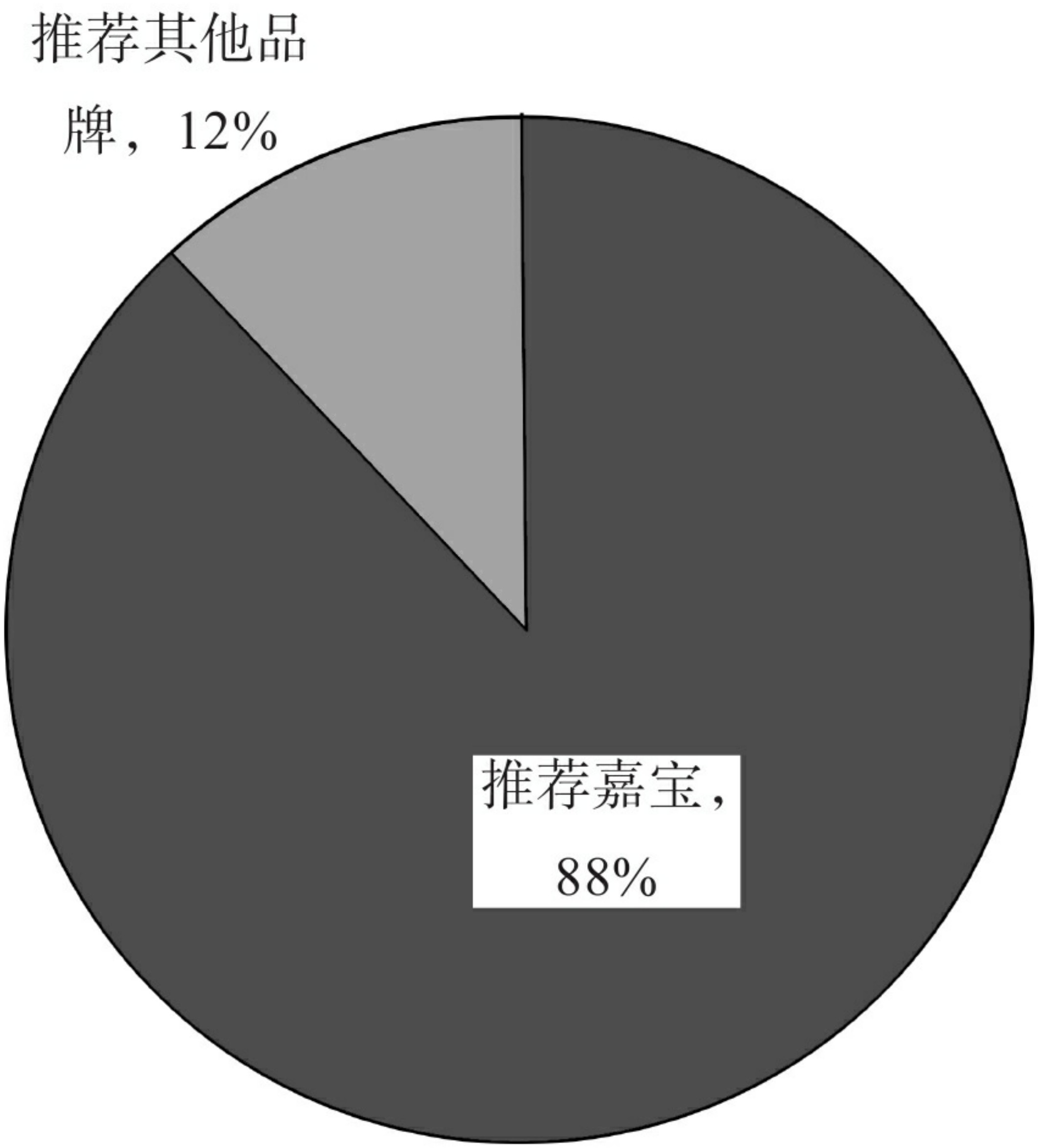


图7－1　推荐某一特定品牌婴儿食品的儿科医生

的确，筛选出来的儿科医生中，有4/5推荐了嘉宝。不过——假如你细看所有数据——那只占了数据总体的12%，同时也只占了经过筛选的样本数据（推荐婴儿食品的儿科医生）的16%。

数据如下：

•嘉宝一开始调查了562名儿科医生。

•408人回应称他们推荐婴儿食品。

•76人推荐某一特定品牌。

•67人推荐嘉宝。

因此，如果你膝下有子女，当你看到或者听到了其中一则广告时，可能会认为4/5的儿科医生推荐嘉宝。即使你留心并注意到数据经过了筛选，你可能认为推荐婴儿食品的儿科医生中，有4/5推荐了嘉宝。但不论你相信那种，结果都是错的。

当时联邦贸易委员会下属单位消费者保护局主管朱迪·伯恩斯坦做了最好的说明：“消费者被引导着去相信嘉宝做了可靠有力的研究并证明了有4/5的医生推荐嘉宝。但是嘉宝完全剔除了不推荐婴儿食品及不推荐特定品牌婴儿食品的医生，肆意篡改研究结果。”㊟

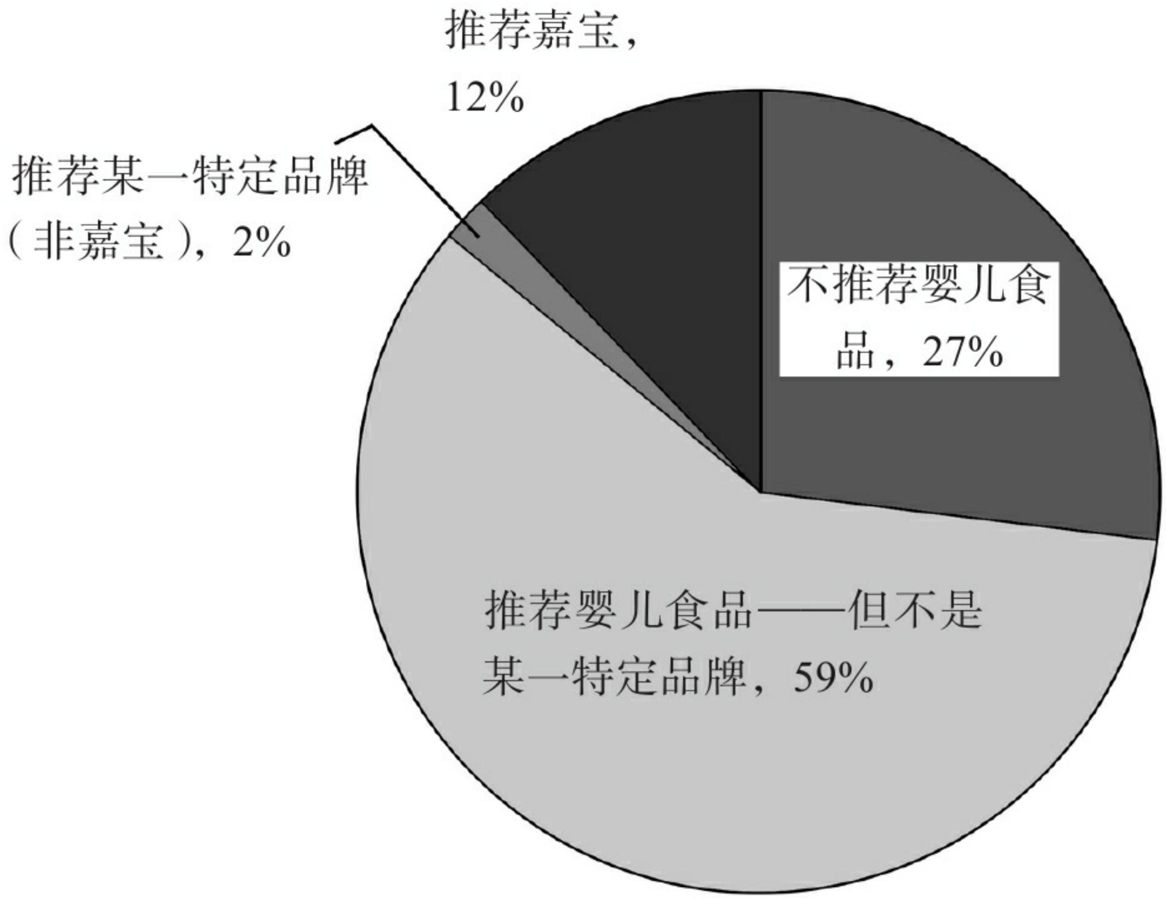


图7－2 所有被调查的儿科医生

联邦贸易委员会迫于公众利益科学中心的请求对嘉宝提出控诉。即使嘉宝明确声明它只谈论推荐婴儿食品的儿科医生的说法，联邦贸易委员会依然认定嘉宝“或明或暗地表示大约有4/5的儿科医生推荐嘉宝”。

换句话说，即使嘉宝告诉人们，它们只将推荐婴儿食品的儿科医生涵盖在内，筛选出最有利于自身的数据，联邦贸易委员会依然认为有些消费者可能会觉得该数据适用于所有儿科医生。联邦贸易委员会认为，即便白纸黑字地写出来，消费者也不能恰当地解读经过筛选的数据。

在一份同意令中——我们应该注意的是，它的宗旨仅仅为了解决问题，不是承认违犯了法律，或者承认所控事实（而非裁决事实）为真实

的——嘉宝同意，以后专业人士对产品任何的“推荐、支持、代言”必须建立在“可靠有力的、可以证实某种说法的科学依据之上”。(注)

---

1. Even if you didn’t want to pay attention, you still had to watch the commercial. DVRs and TiVo weren’t readily available back then, and it was too much of a hassle to tape shows on your VHS player. And if you don’t know what a VHS player is, go ask your parents.
2. Federal Trade Commission, In the Matter of Gerber Products Company, Corporation, Case and Proceedings Docket C-3744, last updated May 30, 1997, https://www.ftc.gov/sites/default/files/documents/cases/1997/05/c3744cmp.pdf.
3. “About the FTC,” Federal Trade Commission website, accessed April 25, 2015, http://www.ftc.gov/ about- ftc: “FTC Accuses Gerber of False and Misleading Advertising,”Federal Trade Commission, March 12, 1997, https://www.ftc.gov/news-events/press-releases/1997/03/ftc-accuses-gerber-false-and-misleading-advertising.
4. Federal Trade Commission, “FTC Accuses Gerber.”
5. Federal Trade Commission, “FTC Accuses Gerber.”
6. Federal Trade Commission, “FTC Accuses Gerber.”
7. Federal Trade Commission, In the Matter of Gerber Products Company Corporation, Agreement Containing Consent Order, File Number 962 3175, accessed July 12, 2015, https://www.ftc.gov/sites/default/files/documents cases/1997/03/gerber.pdf.

筛选数据的"摘樱桃法则"

"摘樱桃"一词源于人们伸手从树上采摘樱桃这一肢体动作。[注]设想一下你在樱桃园里，站在梯子上，手里提着桶，为的是摘下满满一桶能够在市场上销售的樱桃。这样你会略去被碰伤或未成熟的那些，只摘下色泽最佳的装满整个桶。所以，所谓摘樱桃法则，即只选取他人想要的那部分数据（樱桃）。

当然"摘樱桃"还有其他含义，比如只摘长在最低枝条上的水果（打个比方）。例如，打篮球时，"摘樱桃的人"是指游走在对方篮筐附近的球员，等待队友远距离传球，然后他/她就能够很轻易地单手上篮或者灌篮。同样的战术在水球比赛中被称为"海鸥捕食法"，之所以取这个名字，可能因为海鸥会寻找最简单的途径来获取食物。

不过，从本书的目的出发，我们会从筛选数据这一角度探讨摘樱桃法则——同时忽略与期待结果背道而驰的其他数据——从而说明我们的观点。或者，如一项研究报告说的那样，"取其精华去其糟粕"。[注]

那么，为什么说你们要对摘樱桃法则多留个心眼儿呢？我们在第二章就已经讲过，如果你研究相同数据的不同样本，可能会得到截然不同的结果。因此，若你存心这么做，那大可以采取一些手段，得出你想要的结果。

假设有一家小餐馆，从中午到晚上都营业。某一天，餐馆老板每逢整点就数一下就餐人数，得出以下数据：

上午11点——2人

中午12点——25人

下午1点——30人

下午2点——3人

下午3点——0人

下午4点——2人

下午5点——28人

下午6点——35人

晚上7点——5人

晚上8点——3人

次日，一位有意盘下这家店的人来到店里，想了解一下生意怎样。“非常棒，”老板说道，“中午及晚上时间全部满员——队都排到门外面去了！”那人离开了，准备出高价盘下这家小餐馆。

过了一会儿，餐馆的房东走过来，告诉餐馆老板她要涨一倍房租。“您行行好吧！”老板说，“餐馆大部分时间连个人影都看不到！”

当然，两种说法百分之百都是真的——但也百分之百都对数据做了筛选。如果你只看中午、晚上就餐高峰期的数据，餐馆座无虚席；若你只看中午和晚上之间的时段，餐馆几乎空无一人。同样的数据——筛选出的不同的样本——也就带来了极其不同的结果。

---

1. “What Does ‘Cherry Picking’ Mean?” Wisegeek website, accessed August 11,2015, http://www.wisegeek.com/ what- does- cherry- picking- mean.htm.
2. Edward J. Fox and Stephen J. Hoch, “ Cherry- Picking,” Journal of Marketing 69,no. 1 (2005): 46–62.

用统计学的观点如何看祸不单行？

日常生活中我们想到、看到的许多既定概念其实根本不存在，仅仅是老套的数据筛选结果而已。

想想“好事逢三来”这种概念，或者是“坏事逢三来”，或是“好人没好报逢三来”。果真如此吗？还是你对数据进行了筛选？

2009年6月，迈克尔·杰克逊、法拉·福赛特（Farrah Fawcett）、艾德·麦克马洪（Ed McMahon）在同一周之内相继去世。这只是三位名人一起去世的一个例子，是吧？

但是几天后去世的比尔·梅斯，还有2009年7月去世的沃特·克朗凯特（Walter Cronkite），还有在1月去世的约翰·厄普代克（John Updike），又如何解释呢？

美国广播公司新闻发布了一位数学教授的评论，“要死就死三人这类说法纯属无稽之谈，而且在至少两个层面上是毫无意义、可以随意解读的”[注]。“不仅时间范围不确定，而且对于名人也没有明确的含义界定。”

换句话说，如果你想说，杰克逊、福赛特、麦克马洪之死证明了名人要么不死，要么就连续死三人。你得首先把样本集限定在2009年6月23日到6月25日之间，其次断定谁是名人谁不是名人。我们仅仅在谈论加利福尼亚州的名人吗，还是整个美国的名人，为什么不是全世界的呢？我们看到，维基百科上的2009年逝世人物列表中并没有将艾德·麦克马洪列入其中（教训：不要人云亦云），但是同在2009年6月去世的一名日本职业摔跤选手和一位德裔英国社会理论家政治家却赫然在列。[注]

因此，如果你想找到三位名人几乎死于同一时期的例子，是可以实现的，而如果你想找到两位名人死于同一时期的例子，也完全能够办到，甚至四位、五位……随你想要几位都行。这都取决于你如何定义名人，及观察的时间范围。换句话说，它取决于你如何抽取（筛选）数据。

记住：如果你在任意选择你想要的数据来支撑自身观点，那你可能就是在筛选数据。

---

1. John Allen Paulos, “Why Do We Believe That Catastrophes Come in Threes?,”ABC News website, July 5, 2009, http://abcnews.go.com/Technology/Whos Counting/story?id=7988416.
2. “2009: Deaths,” Wikipedia website, accessed April 25, 2015, http://en.wikipedia.org/wiki/2009#Deaths. And no, we don’t recommend Wikipedia as a primary source.

如何解读体育赛事中的统计数据？

如果你是个体育迷，你兴许见过、听过比常人更多的筛选数据的情况。试想：在1988年，你正在观看一场世界职业棒球大赛的比赛——奥克兰运动家队对阵洛杉矶道奇队。第九局下半场，双方同时出局。我们刚看到柯克·吉布森（Kirk Gibson）打出一个全垒打，这个球的精彩程度可以载入史册，仅几分钟的时间，电视屏幕上出现了如下数据："这支首场失利的球队在过去10年间，参加了10届世界职业棒球大赛，7次获胜。"(注)

数据筛选是如何在这里进行的呢？首先，解说员仅仅选择提及在大赛中获胜或者失利一次的球队。我们认为，考虑到比赛正在直播，这可以理解。但是为什么他们将样本集限定在过去10年内呢？为什么不是过去5年内，不是20年内，不是30年内？如果样本集不同，那么所呈现的数据，依然会是同一回事吗？(注)

下次你在看比赛、收听比赛广播，或者读报纸体育版时，留意一下经过筛选的数据类型。这些类型的数据有趣吗？当然有趣。它们有助于人们在看球场上运动员投球、移动、传球时消磨时间。话是没错，可是这些数据究竟有用吗？并不尽然。很多时候，它们比柯克·吉布森在1988年历史性的一天，因为紧张而颤抖的双腿还要不稳定。

好了，那么我们明白了，在棒球比赛中筛选数据不会对任何事产生多少影响。但是，如果那些经过筛选的数据，会花去你10万美元，结果又会如何？

---

1. And, in case you're wondering, the three teams that won game one and went on to win the World Series were the '87 Twins, '84 Tigers, and the '80 Phillies. "1988 World Series- Game 1- Bottom of the 9th," Dailymotion website, accessed April 25, 2015,

http://www.dailymotion.com/video/ xd2fhk_1988- world- series- game-1- bottom- of_sport.

2. It would not. While the team that lost game one went on to win the World Series 7 out of 10 years, that only happened in 3 out of the past 5 years, 11 out of the past 20 years, and 14 out of 30 (using 1988 as our reference year for all). That said, the broadcasters could have chosen an even more dramatic statistic by saying it happened 5 out of the past 7 years.

房价上涨了还是下跌了？

通常你做出的最大的投资就是买房了。因此，你才会在做决定的时候，不愿依赖经过筛选的数据。

假设现在是2012年，你和你的爱人住够了狭小的公寓，认为你们需要买一座房子。周日在公婆家聚餐时，你提起了买房的事，你的公公把快送到嘴边的火鸡放下，然后说这可能是你做得最糟糕的投资。那可不一定，你礼貌地回应道（毕竟谁愿意和公公争执呢？），你指出，房价其实上涨了大约180%。

那么，谁说得对呢？你们都对——这是件好事，因为没有人愿意睡沙发。

你公公认为，房价在过去5年间大幅下跌，图7–3就是他用于证明其说法的房价走势图，看起来似乎无可争辩，不是吗？

但是你公公只看了过去5年的数据。图7–4是你想到的房价走势图，用长远的目光审视了过去十年内的数据。

房价确实略有下降。不过从长远来看，自1982年以来，房价实际上上涨了180多个百分点。

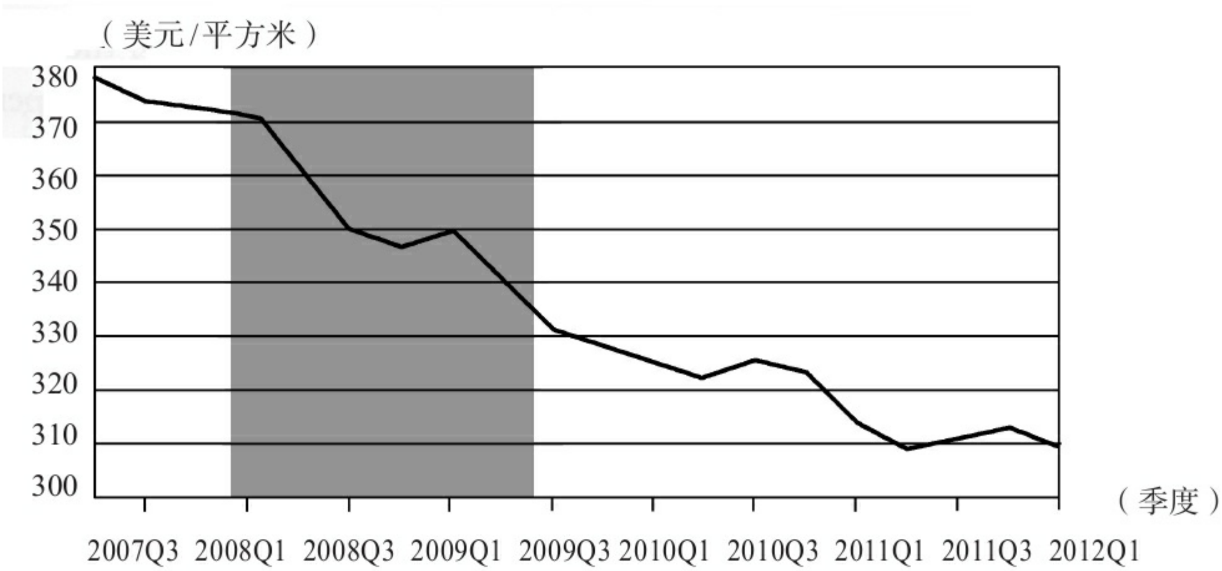


图7－3 美国联邦住房金融局，美国全部交易房价指数（USSTHPI）来自FRED，圣路易斯联邦储备银行，2015年8月25日（https：//research.stlouisfed.org/fred2/series/USSTHPI）

来源：美国联邦住房金融局　research,stlouisfed.org（美元/平方米）

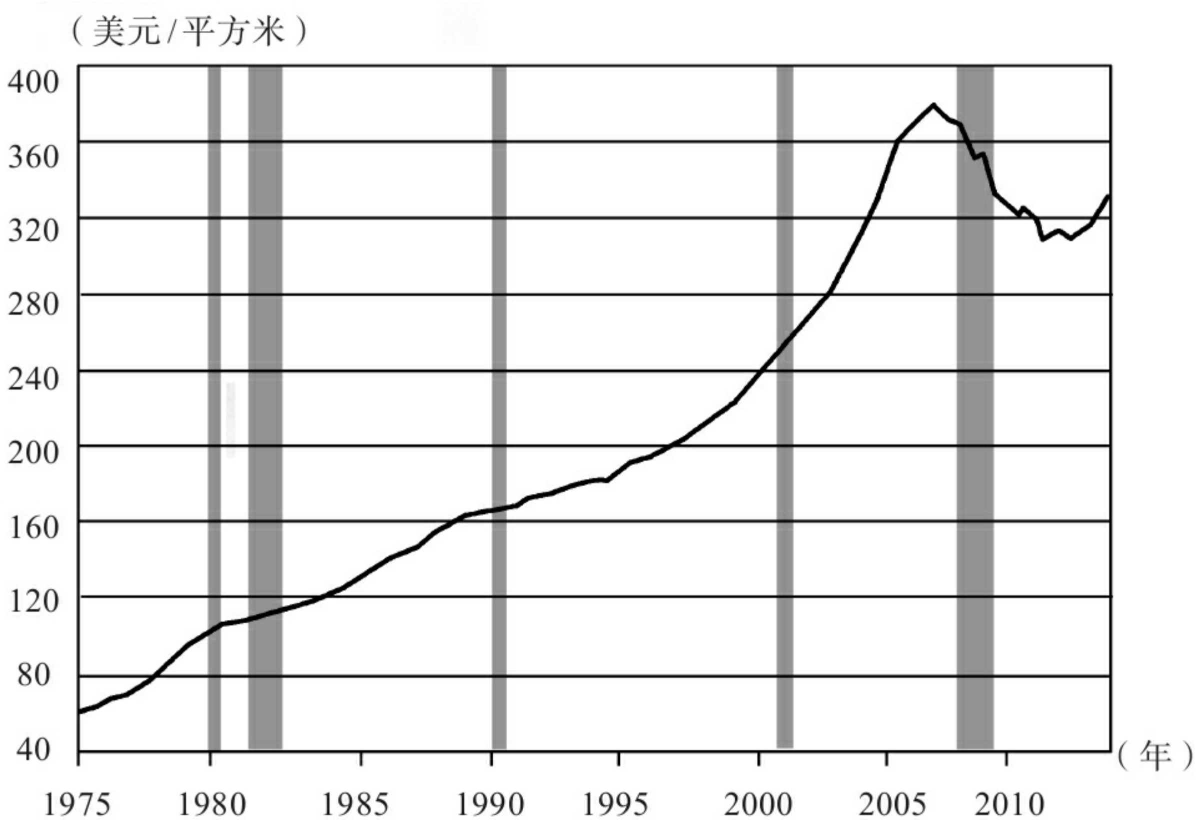


图7－4 美国联邦住房金融局，美国全部交易房价指数（USSTHPI）来自FRED，圣路易斯联邦储备银行，2015年8月25日
（https：//research.stlouisfed.org/fred2/series/USSTHPI）

来源：美国联邦住房金融局 2014 research.stlouisfed.org

就餐期间的争论可能信息量很大，也许你觉得你应当完胜你的公公。然而面对更高风险的讨论，会怎样呢？——比如讨论全球气候变暖？

怀疑科学网站上（The skeptical Science）发布了一张图表，说明你用完全相同的数据来证明地球表面正在变暖——或者证明没有变暖——得出何种结论取决于你如何观察数据。(注)例如，你若看的是大约45年间的数据，气温变化显然有上升的趋势。

不过，如果你采用一模一样的数据，但研究的时间段较短（以8年

或10年为期），你能够轻易地绘制出这段时间内气温略有走低的图表。

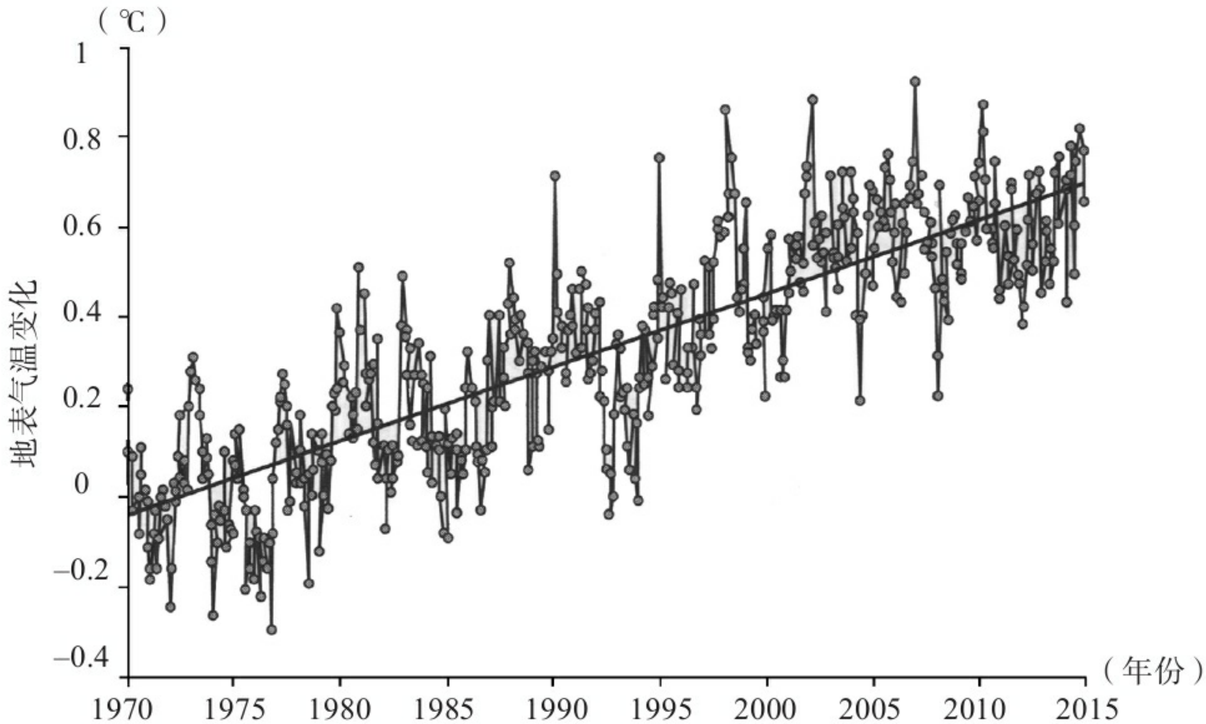


图7－5 德纳·纽斯特里为怀疑科学网所绘现实主义者看待全球气候变暖的方式（SkepticalScience.com）

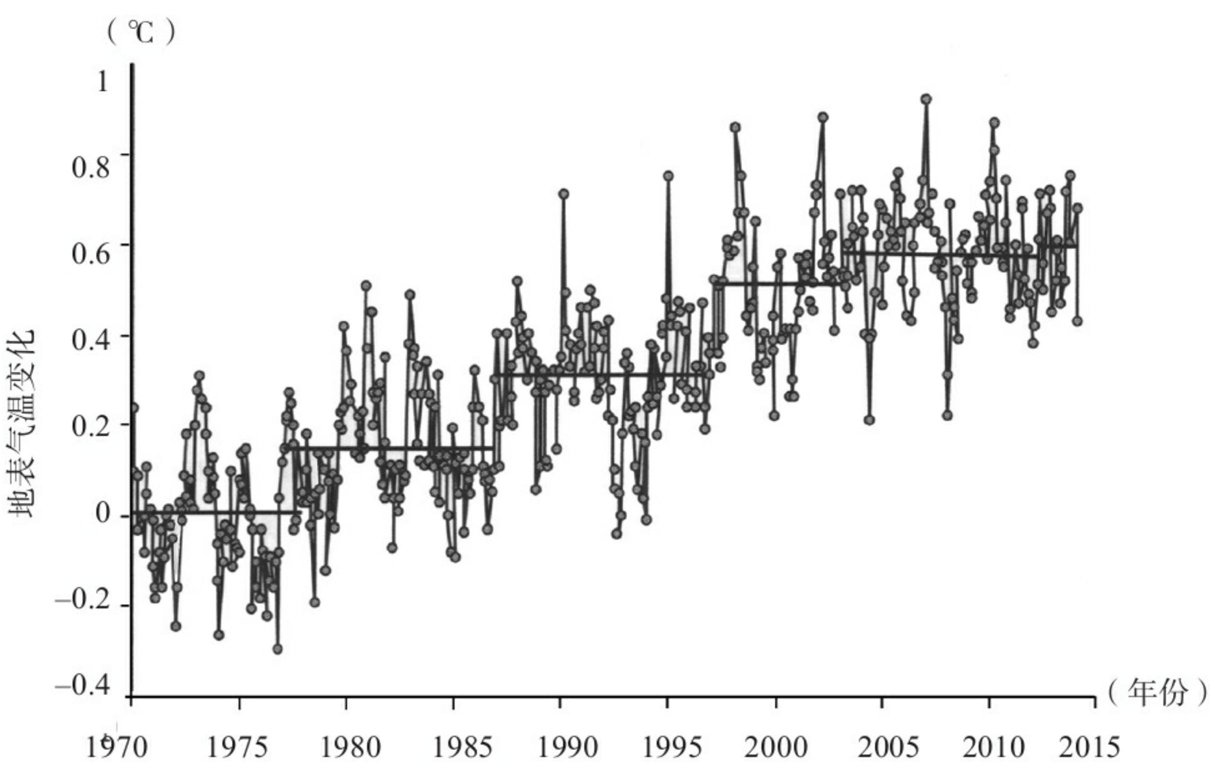


图7－6 德纳·纽斯特里为怀疑科学所绘现实主义者看待全球气候变暖的方式（SkepticalScience.com）

同样的数据，结论却截然相反，一切都归因于数据经过了筛选。

---

1. “The Escalator,” Skeptical Science website, accessed September 2, 2015, http://www.skepticalscience.com/graphics.php?g=47.

政客们如何筛选数据？

好，前文我们谈论了和最值得信赖的人群（儿科医生）有关的内容。接下来，我们来谈谈最不可信的人群。

政客们是筛选数据的专家。如果你在竞选公职，并在步枪协会（NRA）筹款活动上发表演说，你会强调成长过程中与父亲在肯塔基州荒郊野外狩猎的经历。但你若在好莱坞发起募捐活动，你则会谈起在演艺圈大红大紫的舅舅。

在谈及你的竞选对手时，筛选数据的做法跟国旗领针一样常见。在政治舞台上，因为竞选对手精心挑选出来的负面说法、选票、照片等落败的人不计其数。

下面就有一个例子。在1988年的总统竞选中，由于杜卡基斯准许强奸谋杀犯威利·霍顿（Willie Horton）周末出狱休假，副总统乔治·H.W.布什影射马萨诸塞州州长迈克尔·杜卡基斯（Michael Dukakis）对待罪犯心慈手软。一则商业广告充分利用了这一事实，称“迈克尔·杜卡基斯不仅反对死刑，还准许一级谋杀犯离开监狱享受周末”。[注]电视广告成为这场竞选的转折点，帮助副总统布什击败了杜卡基斯。

正如一位记者指出的，如果电视机前的观众了解到“出狱休假的1.1万名囚犯中，99.5%的人都按时返回监狱”的事实，[注]也许竞选结果会大有不同。显然威利·霍顿是未返回监狱的0.5%人中的一员。

数据筛选还有一种方式——断章取义。这里就有一个例子：2003年11月28日，康多莉扎·赖斯——时任乔治·W.布什总统的国家安全顾问——现身哥伦比亚广播公司的《晨间秀》。

如果你去看电影《华氏9·11》（Fabrenbeit 9/11），你会听到一段台词引自赖斯女士。她说：“哦，伊拉克与‘9·11’事件确实有一定联系。”接着摄像机画面就切换了。[注]

完整的论述是这样的：“哦，伊拉克与‘9·11’事件确实有一定联系。不是说萨达姆·侯赛因及其政权莫名其妙地就卷进‘9·11’事件，而是假如你仔细想想‘9·11’事件发生的来龙去脉，正是崛起的仇恨意识形态驱使人们驾机撞击纽约五角大楼。”[注]这与第一句表述的是完全不同的意思。

很明显，赖斯在电影中说的话被从原来的语境中抽离了出来，经过了筛选。《华氏9·11》是由电影导演兼活动家迈克尔·摩尔制作。他承认制作该电影的目的是为了转变人们的想法，影响即将到来的总统选举。[注]我们是否应该料想到一个有目的、有计划的人会对数据进行筛选？当我们走进影院，拿起一本书或者仅仅和某人寒暄几句的时候，又如何获悉这个人的动机是什么呢？

在很多时候，我们做不到——不停问问题就是为了解真相，发掘尽可能多的隐藏数据。

---

1. “Willie Horton 1988 Attack Ad,” YouTube, accessed April 25, 2015, http://www.youtube.com/watch?v=Io9KMSSEZ0Y.
2. Kenneth J. Cooper, “Bush Is Using Case of a Murderer to Assail Dukakis As Soft on Crime,” Philly.com website, June 26, 1988, http://articles.philly.com/1988-06-26/news/26265230_1_michael-s- dukakis- furlough- program- massachusetts- gov.
3. Fahrenheit 9/11, directed by Michael Moore (Culver City, CA; Lions Gate Films,2004.)
4. Brent Baker, “CBS Scolds Bush for Hiding How ‘Kerry Talks Tough’ on Terrorism,” Media Reseach Center website, October 20, 2004, http://www.mrc.org/biasalerts/ cbs- scolds- bush- hiding- how- kerry- talks- tough- terrorism- 10202004.
5. Martin Kasindorf and Judy Keen, “‘Fahrenheit 9/11’: Will It Change Any Voter’s Mind?” USA Today website, June 24, 2004, http://usatoday30.usatoday.com/news/politicselections/nation/president/2004-06-24- fahrenheit-

cover_x.htm.

相信我们——我们在打广告

那产品和品牌为何会筛选数据呢？因为这么做能使它们脱颖而出，脱颖而出才是关键。比如，在我们浏览高露洁的网站时，我们看到31种不同类型的牙膏。注因为你只有32颗牙齿，牙膏看起来似乎太多了。我们到本地杂货店去看一下摆放牙膏的货架，就会看到高露洁旗下就有几十种牙膏——更不用说佳洁士、爱慕（Aim）、艾禾美（Arm & Hammer）、缅因州汤姆家（Tom’s of Maine，高露洁棕榄旗下产品）及其他品牌了。

正如我们在嘉宝一例中所看到的，广告中非常容易发现经过筛选的数据。为什么？因为筛选数据这一做法在广告中大有用武之地，广告商正是凭借这一点来卖产品的。

在你观看电视商业广告的时候，通常来说，你所看到、听到的一切都是经过精挑细选的。无一例外。演员是根据他们的年龄、性别、族裔、体重、发型、雀斑数目等条件精心筛选出来的。尤其是台词，每个字至少要经过12人审查。仔细观察，然后你可能会发现一片布料或者一根柱子也是挑选出来的，因为它们和logo同色。一则商业广告中的事物的存在都不是偶然——这就是为什么拍摄一则30秒的广告可以很轻易地用去12个小时，自不必说剪辑，那会花上数天（甚至数周）的时间。

但在你开口指责广告商仅仅为了卖给你一双新鞋或者一种新型早餐麦片或者人寿保险就用掉这么多时间之前，照一照镜子吧。

说真的，照一照镜子。每次你为约会或者面试穿衣打扮时，你就是在对衣柜里的衣服进行筛选，从而选出最合适的一件。你的约会对象来见你的时候，你是不是要对他说：“哇哦，你故意选这身打扮，选得真

好！”你应该不会这么说吧——至少如果你还想和对方约会的话，你不会这么说。不过这仅仅是我们在日常生活中不停对数据进行筛选的另一个例子而已。注

---

1. “Toothpastes,” Colgate website, accessed July 12, 2015, http://www.colgate.com/en/us/oc/products/toothpaste.
2. Of course, cherry picking should not be confused with curating, which we typically think of as working on behalf of the audience, filtering the data to select the most appealing/interesting data (as opposed to cherry pickers, who often deliberately select only some of the data in order to produce a specific result)

你是“摘樱桃的人”——没错，就是你

现在，你应该可以多少想到几种在日常生活中筛选数据的方式。比如：

•填写线上交友资料时。

•为同事写推荐信。

•不告诉爱人周末去了拉斯韦加斯。

•因为下雨，决定不修剪草坪。

在以上所有例子中，你也许不会将自己掌握的数据全用上，而只采用自己想用的数据。

很多情况下，想要涵盖所有数据几乎是不可能的。想一下你的简历或领英招聘网站的资料就能得出这一结论。根据定义来讲，将你所做事情进行概括，很多时候目的是帮你找到一份更好的工作。因此你当然会将职业生涯中的亮点列举上去，而且只列举那些能够为你增光添彩的资料。这就是筛选数据。你自己知道，你的老板知道，人力资源部经理肯定也知道。这就是设置面试的原因，即发掘未被你列入简历的隐藏数据。

选出最好的，留下其他的

如你所见，你筛选数据的方式有很多。在涉及下列情况时，你可以“选出最好的，留下其他的”：

•日期或者倍数（5年间的房价与10年间的房价对比）。

•群组或子群组（5名推荐婴儿食品的儿科医生中有4名推荐嘉宝）。

•以往结论（“纵观世界职业棒球大赛的历史，曾经以至少6分的成绩赢得首场比赛的15个球队中，夺冠的仅有8个”[注]）。

•含糊或者模棱两可的标签（当你说名人总是三人结伴离世时，那么你会将谁归入“名人”之列呢？）。

•不能公之于众但又非随机选取的因素（谷歌、必应等搜索结果）。

•人（就投票偏好调查某个州亲“右”或者亲“左”的选民得出的结果要比随意调查某个州1000名选民得出的结果差异更明显）。

但是，为了弄清楚数据是否已被筛选过，你得知道存在多少原始数据——话题就转到了第二章所谈及的抽样问题了。如果你在研究气候变化，你有每个大洲过去1000年间每一秒钟的气温材料吗？当然没有。大多数情况下，你掌握不了所有数据的可能性比较大。或许没有人能做得到——或许一些人做到了，但是你呢，只能看到其中的一些数据。

在你思考筛选数据时，要留意离群数据——那些无法让我们对研究对象产生更深理解的特殊数据。你可能认识一些声称能够通过对股票、

基金、行业进行筛选，从而“打败市场”的人。这些人可能仅仅是异类而已——即使是职业的公共基金经理的表现也一直低于市场水平。[注]

筛选数据并不是唯一一项通过忽视整体数据，从而对结论为所欲为的数据策略。你还需注意：

•根据传言提出观点——如果你只依赖传言（也就是说，选取关于某事的故事），那么你就掌握不了全部数据。例如，如果你听说邻居因为在某家餐馆吃饭导致身体不适，并不能断定所有在那家餐馆就餐的人都会身体不适。[注]

•注意力偏向——这种情况发生在你对某些数据尤为关心的时候。注意力偏向与筛选数据的区别在于，筛选数据是特意选取哪些数据该保留，哪些数据该忽略。而注意力偏向可能会让你无法注意到所有可用数据。

•实证性偏见——正如我们在第四章中提到的，这就是找寻支撑你看法的数据趋势。如果你读到一条有关某家餐馆的差评——但是你的老板不管怎样都要拽着你去那里吃饭，自然，你可能会倾向于寻找更多不喜欢这家餐馆的原因。[注]

---

1. Andrew Simon, “Game 1 Rout Not Necessarily Precursor to Title,” MLB.com website, October 23, 2013, http://m.mlb.com/news/article/63287950.

2. Burton G. Malkiel, “Returns from Investing in Equity Mutual Funds, 1971–1991,” Journal of Finance 50 (1995), 549– 572.

3. Not to mention the distinction between causation and correlation, which we talked about in chapter 4.

4. Esteemed economist Daniel Kahneman shared the Nobel Prize in 2002 for his work related to psychological factors that affect our decisions. Much of Kahneman’s work was done in collaboration with Amos Tversky, who passed away in 1996 and was therefore ineligible for the Nobel Prize.

市场营销人员如何筛选数据？

最后，我们来看一下市场营销人员是如何运用筛选数据的方式，使得普通消费者难以看出品牌差异的。

思考一下三种表述，分别来自三家不同的医疗机构：

______“已经第22次位列全国第一了。”

_______“在‘特色医院’评选中，位列全国第一。”

________“医院获得‘美国顶尖医生’称号的医生数量全国第一。”

通过阅读以上表述，三家机构都想让你认为它们是最好的。但假设第一名不允许存在平局，不管是通过什么调查或者研究，那么只有一个可以位居榜首。

那么其他人做了什么呢？他们筛选数据，从而尽可能地强化自己的观点。[注]

---

1. In case you’re wondering, the first example is the Johns Hopkins Hospital (“The Johns Hopkins Hospital Ranked Among the Top Hospitals in the Nation in 2015,” Johns Hopkins Medicine website, accessed September 1, 2015, http://www.hopkinsmedicine.org/usnews/); the second is Mayo Clinic (Mayo Clinic website homepage, accessed September 1, 2015, http://www.mayoclinic.org/);and the third is New York- Presbyterian (“Awards and Recognition,” New York- Presbyterian website, accessed September 1, 2015, http://nyp.org/about/americas- top- doctors.html).

如何识别经过筛选的数据？

如果筛选数据的情况发生在了你身上，你该如何看出其中端倪呢？你可以从以下5点做起。

1.不论何时，在观察或者使用数据的时候，阅读小字。在嘉宝这个例子当中，很多信息可以从注意事项和免责声明中搜集而来。细读注意事项、上下文及脚注。若引用了研究报告——特别是广告商引用了研究报告，你往往会发现大量的限定词，其中一些是经联邦贸易委员会及其他部门许可的。要做一名合格的数据接收者，请拿起放大镜，阅读附注细则。

2.思考一下，数据是否是以随意（或者非随意）的方式选取出来的——这种选取数据的方式，是否可能会让结果朝着某一个方向偏离。比如，频繁、明显、任意切换镜头的现象在体育广播网时有发生。无论何时听到播音员的声音，“过去19场比赛中……”或者“2002年以来”，他们就是在筛选数据了。使用特定时间点的数据难道事出有因？也许你仅仅是在谈论季后赛，或者分析总统大选，或者看看重大事情（像政策变化）发生之前（后）的前兆。但是你需要思考一下你所看到的数据是否经过了随意筛选，或者其筛选的方式（依照时间，等等）是否另有深意。

3.问问是否缺失了什么数据。在嘉宝例子中，我们看到有一大批完全不推荐婴儿食品的医生，还有一组儿科医生，不推荐任何特别品牌。这些小组都没有在嘉宝提出的“4/5”论断中体现。缺失的数据和呈现的数据同等重要。试想有一个杯子里盛了半杯水。乐观者会告诉你一半是满的；而悲观者则会告诉你一半是空的。两者都对——而且都对数据做了筛选。如果你想要做出更好的决定，就要掌握所有数据。

4.不要赋予任何数据超出其自身的意义。想开一些，不要被单个数字左右。它可能看似很有说服力，但是未必能准确地呈现出事情的全貌，因为它是根据经过筛选的数据所得出的。如果一名波士顿红袜队的球迷告诉自己的儿子，击球手贝比·鲁斯[注]创下了出局1330次的纪录，孩子可能会认为贝比是位差劲的运动员。

5.对数据的来源多打个问号常常大有裨益。比如，我们喜欢喝咖啡。本书有一半内容是在咖啡店完成的（我们还需满满一张致谢，只为所有这些友爱和善的咖啡师）。但如果你读到一篇来自美国国家咖啡协会（NCA）的报告，报告中引用了团队总经理及首席执行官的话，称人们喝咖啡喝得越来越多了，听到这样的话，你不该觉得惊讶。毕竟国家咖啡协会的核心目的是"在咖啡全球化的大环境下，确保美国咖啡业保持领先地位"。[注]随便读一本贸易组织编写的读物，其中必然充斥着称赞其成员企业的文章。（尽管也有例外——《华盛顿邮报》曾经报道过，"有一项得到奶酪行业赞助的研究，得出了奶酪有害健康的结论"。[注]）同往常一样，在你接收数据的时候，请思考一下数据的来源。

因此，下次你的老板说因为"上月销售额下滑"不能给你涨工资时，问问她过去一个季度、一年或者10年的销售额。

当你12岁的孩子因为"她所有的朋友都在看R级电影"而想看R级电影时，问问她数据组中是否涵盖了她的每一个朋友——还是仅仅局限于爸妈对这类事情比较无所谓的朋友呢。

还有如果你听到有人说"4/5的人……"，那么问问题的时候到了。

---

1. 贝比·鲁斯，美国职业棒球运动员，有"棒球之神"美称。——编者注

2. "Press Release: NCA Publishes Annual Consumption Tracking Study," National Coffee Association press release, June 18, 2013, http://www.ncausa.org/custom/headlines/headlinedetails.cfm?id=876&returnto=778. Retrieved 3/15/15.

3. Roberto Ferdman, “A Study About Butter, Funded by the Butter Industry,Found That Butter Is Bad for You,” Washington Post, August 7, 2015, http://www.washingtonpost.com/news/wonkblog/wp/2015/08/07/ the- butter- industry- probably- regrets- paying- for- this- study- that- shows- butter-is- bad- for- you/.

第八章 为什么福岛事故是可以预防的：预测未来的技术

2011年3月11日，日本发生了该国有记录以来最严重的地震。据美国国家航空航天局的科学家称，此次东日本大地震震级达9.0级，震中位于太平洋海域，距离日本沿海不远，其威力之大，使得地球的形状轴（形状轴周围的地球质量是平衡的）位移距离超过0.5英尺（15厘米）。[注]

地震发生不到半个小时，引起的巨大海啸就席卷了日本福岛第一核电站。原本核反应堆一感知到地震就会自动关闭，但是海啸摧毁了应急发电机和冷却泵，还损坏了多级反应器。结果，“数量巨大”的放射性物质随之泄漏，[注]数百万吨被污染的水也涌入海水之中。[注]

在国际核事件分级表中，该事故被评为7级事故——这已经是最高级别了，只有重大事故才可以被评定为这个级别。当时，只有切尔诺贝利事故能与之相提并论。当然，不容忽视的还有地震和海啸在日本全国造成的伤亡，据报道15391人因之丧生，8000多人失踪。[注]

地震——往往会引发巨大的海啸——在日本是人尽皆知的自然灾害。如今，日本的一个国家机构准备了若干幅日本全国地震灾害图，并用颜色代码做了标记，突出全国范围内地震事故发生的概率。[注]甚至在当初策划建设福岛核电站时，工程师也知道面临地震及海啸的风险，所以设计出来的核电站可以经受3.1米高海浪的冲击。这一高度是基于当时的一般惯例，即根据历史上的海啸记录判断未来海啸的浪高。3.1米的浪高是根据1960年侵袭智利沿海的大地震而来的。[注]

实际上，据估计席卷福岛第一核电站的海啸浪高14~15米。由于测量仪在海啸中遭到毁坏，因此我们无从得知海啸的确切浪高。

核电站的确是遭受了自然灾害的袭击，而之后的一系列事故，是由于无法准确地预测灾害的强度，导致无法应对才引发的。正如独立委员会官方报告上写的：“福岛事故是一场本就能够且应当预见，从而避免发生的人为灾难。”[注]

当然还有一些其他因素。据官方报道称，东京电力公司、监管机构及政府被指“相互勾结”。[注]有些计算机模型没有准确计算海啸冲上建筑周围土地情形，在那种情况下，海浪的高度会增加。但根本来讲，设计期间所做的预测在若干年后带来了严重的后果。

最开始对福岛核电站能够承受3.1米海浪袭击的评估来源于核电站设计前几年间的数据记录。不过据国际原子能机构称，在地壳运动高度活跃的地区，可能需要研究“数万年间”的数据。[注]在一篇卡内基国际和平基金会的报告——《为什么说福岛事故是可以预防的》（*Why Fukushima Was Preventable*）一文中，研究者们引用了一篇报告（发表于福岛事故几年前）。引用的报告称，根据沉积物推断，该地区“每隔千年就会被一场大规模海啸席卷”。[注]卡内基报告的另一项研究（发表于2011年海啸后不久）列入了发生在日本及附近的6次海啸，500年来最大的海啸浪高达20多米。[注]《纽约时报》甚至报道，存在“所谓的海啸石，有的已经存在了600多年”，还刻着诸如“勿在此地建设家园”的警告[注]。

在本章，我们会探讨一下预测在日常生活中所扮演的角色，讨论内容涵盖地震、抛硬币、选举、天气等。还会讨论为什么对未来的预测很难做到精准，但也不是说绝无可能。

---

1. Alan Buis, “Japan Quake May Have Shortened Earth Days, Moved Axis,”NASA website,

March 14, 2011, http://www.nasa.gov/topics/earth/features/japanquake/earth20110314.html.

2. The National Diet of Japan, "The Official Report of the Fukushima Nuclear Accident Independent Investigation Commission," 2012, https://www.nirs.org/fukushima/naiic_report.pdf.

3. Patrick J. Kiger, "Fukushima's Radioactive Water Leak: What You Should Know,"August 9, 2013, National Geographic News website, http://news.nationalgeo graphic.com/news/energy/2013/08/ 130807- fukushima- radioactive- water-leak/.

4. International Atomic Energy Agency Mission Report, "The Great East Japan Earthquake Expert Mission, IAEA International Fact Finding Expert Mission of the Fukushima Dai- Ichi NPP Accident Following the Great East Japan Earthquake and Tsunami," May 24, 2011, http:// www- pub.iaea.org/MTCD/meetings/PDFplus/2011/cn200/documentation/ cn200_Final-FukushimaMission_Report.pdf.

5. "J-SHIS Japan Seismic Hazard Information," J-SHIS Map website, accessed July 26, 2015, http://www.j-shis.bosai.go.jp/map/?lang=en. That said, the Guardian noted that, "since 1979, every earthquake that caused 10 or more deaths has struck in regions claimed to be at low risk." (Ian Sample and Justin McCurry,"Flawed Earthquake Predicitions Gave Fukushima a False Sense of Security,"Guardian website, April 13, 2011, http://www.theguardian.com/science/2011/apr/13/ flawed earthquake- predictions- fukushima.)

6. IAEA Mission Report. Based on this data, the site level for key structures and systems was set at 4.0 meters, while the elevation for critical buildings was set at 10 meters. In 2002, Tokyo Electric Power Company (aka TEPCO, the plant operator) revised the estimated height of a tsunami to 5.7 meters. Then, on March 7,2011—just four days before the deadly disaster—TEPCO told Japan's Nuclear and Industrial Safety Agency that a future tsunami could be higher than 10 meters, according to an article in the Asahi Shimbun ("TEPCO Warned of Big Tsunami 4 Days Prior to March 11," Asahi Shimbun website, August 25, 2011,http://ajw.asahi.com/article/0311disaster/quake_tsunami/AJ201108257639).

7. Official Report, National Diet of Japan.

8. Official Report, National Diet of Japan.

9. IAEA Safety Standards Series, "Site Evaluation for Nuclear Installations, Safety Requirements No. NS-R-3," International Atomic Energy Agency, November 2003, http:// www- pub.iaea.org/MTCD/publications/PDF/Pub1177_web.pdf.

10. James M. Acton and Mark Hibbs, "Why Fukushima Was Preventable,"Carnegie Endowment for International Peace, March 2012, http://carnegieen dowment.org/files/fukushima.pdf.

11. Ludger Mohrbach et al., "Earthquake and Tsunami in Japan on March 11,2011 and Consequences for Fukushima and Other Nuclear Power Plants," VGB Power Tech (Germany),

April 1, 2011, www.vgb.org/vgbmultimedia/News/Fukushimav15VGB.pdf.

12. Martin Fackler, “Tsunami Warnings, Written in Stone,” New York Times, April 20, 2011, http://www.nytimes.com/2011/04/21/world/asia/21stones.html?_r=0.

明天太阳会照常升起吗？

预测建立在一个看似简单的问题之上：立足过去，我们能预知未来吗？

只是这个问题一点都不简单。实际上，这是本书最复杂的问题之一，正因为这样，我们才把这个问题留到最后讨论。

我们就以前半个问题展开吧："立足过去……"

如果你拥有所有以往数据，也就是我们所谓的数据总体，那么你就有了一个良好的开端。例如，不管以什么方式，若我们能够观测到曾经发生在日本的每一次地震及海啸，我们就掌握了所有这些事件的数据。即使那时我们常常禁不住思考未来还会发生什么别的事情。

但如果我们没有搜集到全部数据，就不得不依赖某一个样本。而在抽样的过程中，我们也不是总能知道样本是否代表了数据总体的真实情况。所以你得出的结论也不确定，也就是抽样误差——我们在第五章讨论过。这是福岛大灾难背后的因素之一，鉴于大地震及海啸发生的频率相对较低，其样本容量（几十年）太小，不具备代表性。

样本的不确定性越大，预测的不确定性就越大。掌握的信息越详尽，预测结果越准确，在统计学中，我们将预测的基础称为"模型"。模型展示了所有掌握的信息——你用来预测未来的因素，你所依赖的基础样本数据，以及你通过数学方法确定的各因素之间的关系。换句话说，该模型体现了你所认为的各因素之间的关系。毕竟，如果你解释不了已经看到的东西，就很难（几乎不可能）解释在未来将要看到的东西。㊟

以往的数据也会受到目前我们讨论过的许多因素的影响——筛选数

据、遗漏变量、离群值、显著性差异等。以上任何因素在你的以往数据中出现都会导致影响预测的问题，因为未来通常立足于以往的数据。

现在，你一旦开始着眼未来，而不是回首过去，问题就会变得愈加复杂，因为你将存在于过去和未来的不确定性叠加在一起了。过去发生过的事情未必会在未来发生。

比如，你想预测来年小麦的价格。你搜集了人类有史以来所有关于小麦价格的数据，还有决定小麦价格的诸多因素（气温、肥料价格、运输费用等）。首先，你需要建立一个统计模型以确定过去影响小麦价格的因素是什么，以及不同因素之间存在怎样的数学关系。然后，凭借此模型预测来年的小麦价格。㊟

问题是，不管样本数量有多大，你的统计模型有多精确，依然存在导致预测出现偏差的未知因素：

•要是铁路运输费用翻了一番，会怎样？

•要是国会通过了新的法案，限制小麦价格，会怎样？

•要是小麦出现基因突变，成长速度比原来翻了一番，导致全球小麦供应量翻倍，会怎样？

我们不知道以上情况是否会发生，但是我们也不知道它们会不会不发生。但是这类的干扰事项（有时被称为结构性变化）会扰乱我们立足过去精准预测未来的能力。㊟

预测并非易事，因为我们对世界运行方式的认知上还存在盲区。存在不可预知的事情，未知的进程、随机发生的事情。人说变就变，事物也不会一成不变。你在研究的数据可能会改变——你对基础过程的理解同样也会变化。例如，福岛灾难之后，根据日本民事服务协会提出的新

研究方法，海啸的预期浪高由3.1米提高到了5.7米。[注]当你想要预测未来时，这些变化就是你需要铭记在脑海中的注意事项。

---

1. Even the best model will not fully predict the past because there is, at least in theory, truly random error. In fact, although it may seem counterintuitive, a model that perfectly explains the past is often a sign of an “overfit” model and will perform poorly in the future.
2. With complex forecasts— such as determining where a hurricane will make landfall— forecasters may look at a set of models. Even small changes in the model may have an impact on a forecast. For example, according to an article on IEEE Spectrum (Tekla Perry, “Predicting Hurricane Sandy,” accessed August 2, 2015, http://spectrum.ieee.org/ tech-talk/computing/software/ predicting- hurricane- sandy) simply using two different computers to run the same software may result in two different outcomes, since each computer may run calculations in a different order. But if the forecasters make minor changes and still find consistency in the various forecasts, they can rest (somewhat) assured that the models are accurate.
3. For example, building contracts frequently include what are called “force majeure” clauses that specify types of natural disasters that can result in delays in construction.
4. IAEA Mission Report.

预知和预测的区分

地震是无法预知的。至少美国地质调查局（USGS）认为不可能做到，而且他们还为“预知”和“预测”做了明确的区分。他们指出，“不存在可以预知某次地震发生的科学易行的方式”，还补充说“根据人们的概念，所谓预知，需要能够预言出未来地震的震级、发生的时间及地点，这在现在看来，是不可能做到的。”[注]目前我们缺乏准确预知地震的数据和技术。

这就是说，美国地质调查局说，这些地方“从长远来看最有可能发生地震”。只有在评估在某一段时期内地震发生的可能性时，他们才将其称为预测。

那么我们就会面临这样一个问题：预知与预测到底存不存在区别。内特·希尔在《信号与噪声》（*The Signal and the Noise*）一书中指出，有的人（多数是研究地震的知名地震学家）会区别使用这两个词，而有的人则会将两个词互换使用。有的人认为“预知”一词具有双重含义——发生或者不发生——而预测的更多是概率事件，事情发生的概率有百分之多少。（说得复杂点，当谈论到过去、现在或者未来的数据时，可能需要进行估算。）

我们发现在某些情况下，这类区分确实存在。然而，我们从撰写本书的目的出发，一般情况下将“预知”与“预测”看作同义词，如果不是，会另行注明。

为什么呢？

首先，因为这本书的读者群是普通的数据接收者，而不是地震学家或者高级统计学家。

其次，我们不想让字词语义上的区别喧宾夺主，盖过本章主旨——阐释运用数据预测未来时应该或者不应该采用的方式。

---

1. “Earthquake Facts & Earthquake Fantasy,” USGS website, accessed July 26,2015, http://earthquake.usgs.gov/learn/topics/megaqk_facts_fantasy.php.

出国旅游之前，请告知金融机构

你遇到过信用卡公司误认为你的卡被盗刷而给你打电话的情况吗？这时有发生。你在欧洲度假或者在品牌折扣店血拼，突然接到了来自信用卡公司的电话，原因是该公司担心你的卡出现了欺诈行为。

为什么呢？可能是因为公司参照了以往的数据，并以此创建了你未来购买行为的模型。那么当数据和模型不相符的时候，你就会接到电话。［Visa（维萨）在其网站上建议，如果你要出游，请告知金融机构，这样"有助于确保你的卡不会因为异常活动而被标记"。(注)］这就是假阳性（false positive）的一个完美例子——信用卡公司预测，你卡上的消费行为有可能是欺诈行为，然而实际并非如此。

统计模型无法解释说明的事件就是预测误差的潜在之源。正如抽样误差为我们指出了抽样存在的不确定性，预测误差则是衡量未来不确定性的一种途径，基本上的方式是，在误差出现时比较预测结果与现实结果。(注)

往往用预测区间（prediction interval）来衡量预测误差，预测区间就是我们希望从中看到下一个数据点的区间。当信用卡公司因为"被盗刷"的信用卡而致电你的时候，可能是因为你最新的消费记录超出了它的预测区间。(注)因此，致电Visa并告知他们你打算去夏威夷度假，实际上你扩大了Visa的预测区间，连夏威夷海滩上的迈泰鸡尾酒都能囊括在内了。

你可能知道——至少本能地知道，距离要发生的事情时间越近，预测结果越精准。从统计学上讲，就是说你的预测误差和预测区间变小了。假设现在是5月1日，对比次月销售额与本年其余时间销售额的预测

结果，对次月销售额的预测结果的误差（预测区间）可能更小，有以下两个原因：

1.你掌握了更多、更新的历史样本数据。如果你试图在5月1日判断年终销售额情况，你只有当年4个月的数据。如果你等到10月1日，此时你有9个月的数据。在所有条件相同的情况下，历史数据越多，你做出的预测往往就越精准。

2.类似地，从理论上讲，对比从5月1日至6月1日与5月1日至12月31日两个时间段，发生在前一时间段的预见不到的事情要少。时间越长，影响你预测的因素出现的概率越大（新竞争者、不同的管理环境、供应商价格变动，不一而足）。

只是要记住：预测误差跟抽样误差是两回事。在政治民调中，我们会遇到这种困惑，政治民调常常论及误差范围，似乎这仅仅是抽样问题。但是抽取更多的人——甚至全部人口，也不能完全消除预测误差。政治候选人所获的支持会随着时间推移而变化。有的候选人会在其竞选对手退出时获得支持，而有的候选人则因为辩论中的不佳表现或者政治集会上不合时宜的呐喊（“迪安尖叫”[注]）而失去潜在的选民。[注]

最后，我们来谈谈预测偏差（forecast bias），该词被用来描述一贯偏高（正偏差）或者偏低（负偏差）的预测误差。为什么会发生预测偏差呢？也许是你所使用的模型存在误差。比如，你在本地经营一家自来水公司，你想用一种模型来预测年度供水需求，而该模型没有将城镇人口每年增长5000人的事实考虑在内，那么你得到的可能就是负预测偏差——需求总是偏低。当然，预测偏差也可能是蓄意发生的。假设你有一家分公司，次年的预算分配是根据你的预测而做。你预测得越高，分公司得到的钱就越多。如果存在这类动机，弄清预测偏差是如何发生的就轻而易举——如果你的工作需要问人们的预测，你就需要将这点铭记于心了。

---

1. “Travelling with Your Visa Card,” Visa website, accessed August 2, 2015, http://www.visa.ca/en/personal/travelling/index.jsp.
2. Of course, there are some situations that are entirely predictable. For example,if you have a fixed interest rate on a loan, you can calculate exactly what your payments will be for the future.
3. Statisticians may also measure the tolerance interval, which shows how confident we are that our prediction interval contains a certain percentage of the expected values.
4. 迪安尖叫：2004年美国总统候选人迪安在对自己的支持者讲话时大喊一声“耶”，被一些选民认为失态。——编者注
5. “Howard Dean,” YouTube video, 0:23, posted by Kittensaremegasilly, October 22, 2006, https://www.youtube.com/watch?v=KDwODbl3muE.

正确地看待预测

我们采访了伊利诺伊大学教授、商学院商业与公共政策中心主任杰弗里·布朗（Jeffrey Brown），他说："正确看待预测的方式是将其看作一种缩小出现不同结果可能性的手段，即缩小'任何事情都可能会发生'与'这就是可能会发生的结果'二者间范围。"

但问题是——据布朗称（他也是国家经济研究局退休研究中心的联合主管，为决策者提供信息）："政治家与民众不会过多地关注不确定性，他们只想要听到简单扼要的话，诸如'社会保障系统要破产了'或者'社会保障会好起来的'之类。"

任何评估都存在不确定性，然而这并不是有些人所喜闻乐见的——他们只想要数字。

正如布朗所说，"大多数经济学家明白，我们真正在做的是'规划'而非'预测'"。换句话说，如果生育率、死亡率或者劳动参与率以特殊的方式演变的话，我们可以非常自在地将社会保障系统内的资金运转方式转化为模型。但若要明确地说明陈述生育率、死亡率或者劳动参与率将以哪种特定的方式发展，我们就没那么自在了。

不过除了在数据统计方面受过训练（或者读过这本书）的人，这类的细微差别可能对他们来说无足轻重。布朗指出："结果是，我们常常见到这样的情形，当政策没有完全达到预期效果，哪怕结果仍处在初步估算出来的置信区间时，人们就已经表达出意外、失望乃至愤怒之情了。"

抛硬币与“赌徒谬论”

到目前为止，我们谈论到的许多预测都是决定性的：你从预测中得到一个精确的结果。比如：“今天要下雨。”这跟地震学家“预测”地震的方式类似（旧金山3月2日将会发生3.1级的地震）。

判断抛出的硬币正面朝上还是反面朝上就是可能性预测（probabilistic forecast）一个很经典的例子。可能性预测即你能够推断出结果发生的概率。比如：“今天下雨的概率为20%。”这跟地震学家“预测”地震的方式类似。

当你抛硬币的时候，你知道要么正面朝上要么反面朝上。当然，它也有可能立起来。不过鉴于有文章估算说立起来的可能性为1/6000，我们就暂且忽略不计，为了本节主旨，假设落下的硬币只会正面或者反面朝上。[注]正如《美国统计学家期刊》中的一篇文章所指出的，我们还假设“抛硬币不可能出现误差”，这样硬币落下的时候极有可能是正面或者反面朝上。[注]而且我们是抛硬币，而不是旋转硬币——比如，一篇《在线科学新闻》上文章称，一枚旋转的硬币反面朝上的概率达80%，这是硬币的重心转移到正面的缘故。[注]）

因此，通过我们用抛出的、不会产生重心转移的、不会立起的硬币，可以得出以下结论：

硬币落下时正面朝上的概率是50%

硬币落下时反面朝上的概率是50%。

如果你抛硬币抛了两次，那么一次正面朝上，一次反面朝上的概率为50%。

从现在开始，事情就会变得越来越有趣了。

如果你抛10次，硬币5次正面朝上5次反面朝上的概率不是50%，实际概率甚至低于25%。

如果你抛100次，硬币50次正面50次反面朝上的概率大概为8%。

哈？

你们从中可以看出的是，（理论上）应该发生的事情与现实中实际发生的事情是有区别的。硬币本身并不知道它落下时应该有一半正面概率朝上另一半概率反面朝上。这是关于预测重要的启示——因为最有可能发生的事情并不意味着真的会变成现实。

许多人也因抛硬币而困惑不解，因为他们认为，比方说，如果你抛两次硬币，那么硬币两次都正面朝上的概率为1/3，都反面朝上的概率为1/3，一正一反的概率为1/3。但是由于得到一正一反结果的方式有两种——A硬币正面朝上，B硬币反面朝上或者A硬币反面朝上，B硬币正面朝上——那么实际上硬币两次都正面朝上的概率为1/4，都反面朝上的概率为1/4，一正一反的概率为1/2。记住，在你做预测的时候，要将所有可能发生的结果都考虑在内。

当然，你抛硬币的次数越多，你的结果就越接近于“五五开”。为什么呢？因为你抛硬币的次数越多，你实际上是从无穷尽的抛掷结果中选取了更大的样本容量。那么较之于小样本容量，较大的样本容量——假设其他条件相同——将给予你更精准的结果。尽管大样本容量不是必要条件。1936年，《文摘杂志》（*Literary Digest*）在一项调查中收到了230万个回复，并预测阿尔夫·兰登（Alfred Landon）在美国总统竞选中将得到57%的选票，从而击败富兰克林·D.罗斯福。而实际上，罗斯福得到了60%多的选票。

那么，我们谈过用以往数据预测未来，还记得吗？而在很多例子中以往数据无足轻重，抛硬币这个例子就是其中之一。那是因为我们知道抛硬币（至少从我们叙述方式上看）是一件随机的事情，而你不能将过去作为模型用于随机事件当中。

如果你抛10次硬币，硬币每次都正面朝上，那么硬币下次正面朝上的概率仍然为50%，即使你抛100次、1000次，或者100万次，硬币次次都正面朝上，下次硬币正面朝上的概率仍然是50%，因为我们掌握了已知的模型，在此模型中，硬币正面朝上、反面朝上的概率均为50%。

很多人弄不明白。他们直觉上认为下一次必须是反面朝上。而从统计学上看，他们大错特错。的确，你不可能连续10次抛出正面。但是这依然跟未来没有丝毫关系。“五五开”事件随着时间的推移会发生，但是每次抛硬币都是独立事件，与过去发生的事情毫不相干。(注)

人们的这种直觉被称为“赌徒谬论”——认为你输一阵之后，就会赢了。根据得州农工大学（Texas A&M）做的研究来看，“实际上，你赢的概率跟先前并无区别”。(注)（好消息是，错不在你。研究者构建了一个脑内神经元模型，发现“倾向于改变诸如正面–反面模式的神经元的数量远远多于倾向重复正面–正面的神经元的数量”。这个结论真是好。但要记住，得出这个结论的基础是神经元模型，而非真实的神经元。）

作为明智的数据接收者，在分析预测时，你需要留心赌徒谬论。扪心自问：该预测是建立在以往数据发挥重要作用的模型上吗？或者预测的是完全独立于过去的事件？(注)

《乐透统计数据》（*Lotto Stats*）杂志上有一句话，我们在这里摘下来分享给大家，话是这么说的：“当你在玩‘每日数字游戏’的时候，面前的信息越多，赢的概率就越大。”(注)

捂上耳朵吧，孩子，因为说这句话的人是在睁眼说瞎话。

1. Daniel B. Murray and Scott W. Teare, "Probability of a Tossed Coin Landing on Edge," Phys. Rev. E 48, 2547 (October 1, 1993), http://journals.aps.org/pre/abstract/10.1103/PhysRevE.48.2547.
2. Andrew Gelman and Deborah Nolan, "You Can Load a Die, But You Can't Bias a Coin," American Statistician 56 (November 4, 2002), http://www.stat.columbia.edu/~gelman/research/published/diceRev2.pdf.
3. Erica Klarreich, "Toss Out the Toss-Up: Bias in Heads-or-Tails," Science News Online, 165, no. 9 (February 28, 2004): 131, http://web.archive.org/web/20080314023237/http://www.sciencenews.org/articles/20040228/fob2
4. Just make sure what you're looking at is truly an independent event. For example, if you're a basketball player shooting free throws, each throw is not an independent event. Yes, it's a separate event from the one before. But it's also influenced by other, past factors— everything from your fatigue to your state of mind after making (or missing) your past shot. That said, there is conflicting research in terms of whether or not athletes can have a s-called "hot hand," with one paper finding that " [t] he belief in the hot hand and the 'detetion' of streaks in random sequences is attributed to a general misconception of chance" (Thomas Gilovich, Robert Vallone, and Amos Tversky, "The Hot Hand in Basketball: On the Misperception of Random Sequences," Cognitive Psychology 17 (1985): 295– 314, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6700), while a study from Harvard found that "players who are outperforming will continue to do so, conditional on the difficulty of their present shot" (Andrew Bocskocsky, John Ezekowitz, and Carolyn Stein, "The Hot Hand: A New Approach to an Old 'Fallacy,' " presented at the MIT Sloan Sports Analytics Conference, February 28–March 1, 2014,http://www.sloansportsconference.com/wp-content/uploads/2014/02/ 2014_SSAC_The- Hot- Hand-A- New- Approach.pdf).
5. Ellen Davis, "Committing the 'Gambler's Fallacy' May Be in the Cards, New Research Shows," Texas A&M Health Science Center website, March 9, 2015,http://news.tamhsc.edu/?post= committing- the- gamblers- fallacy- may-be-in- the- cards- new- research- shows. Thanks to Ron Friedman for the find.
6. There's another way of looking at this, known as Bayesian probability (after the eighteenth-century English mathematician Thomas Bayes). With Bayesian probability, you use the data gathered to update your initial beliefs after the fact.It's the opposite of the way in which the gambler's fallacy works. As one of John's colleagues pointed out, it's the difference between knowing that a coin is fair and learning about the coin. So, a Bayesian might flip a coin 10 times, get heads all 10 times, and adjust his probability to say that the coin was always more likely to land heads up. Here's another way to think about it— consider a bowl full of M&M's on top of a shelf. You can reach inside the bowl, but not see inside it.Your initial belief is that

approximately 24 percent of the M&M's are blue (the colors aren't equally distributed, according to a blog post from Josh Madison—https://joshmadison.com/2007/12/02/ mms- color- distribution- analysis/— and other research). You reach inside, take out an M&M, observe the color (blue),and then put it back. You do this 10 times, and each time you get a blue M&M.For non- Bayesians, the probability of getting a blue M&M on your next turn is still 24 percent. But for a Bayesian, the probability is higher, since you've now revised your initial beliefs (based on the sample) to believe that more than 24 percent of the M&M's are blue.

7. "3 & 4 Digit, New York's Best Bi-Weekly Numbers Guide," Lotto Stats website,June 18, 2008, https://www.lotstats.com/pdf/lssampleissue.pdf. That said, if your goal is to split the jackpot with as few people as possible, there are strategies you can employ. For example, you can pick less popular numbers, such as those above 31 (since many people use dates as their "lucky" numbers), and you can choose consecutive strings of numbers, since many people think those combinations are less likely to win. Just don't forget about your favorite book authors when you hit it big.

我们身边的预言家

从诺斯特拉达姆斯（Nostradamus）据称成功预言了伦敦大火，到当今社会想要成为预言家的人断言“末日不远了”，世界各地的人长久以来都声称自己具有预测未来的能力。甚至在舞台上，预言家也扮演着很关键的角色，他们警告俄狄浦斯会杀死自己的父亲（然后同母亲结婚），还恳求尤利乌斯·恺撒“在3月15日（Ides of March）这一天要当心”。

今天，仍然存在依靠准确预测未来的能力谋生的男男女女。

不过他们不是预言家，不是先知，也不是神使。

在这个案例中，他们是对冲基金经理。

对冲基金是从不同投资者募集资金的投资项目。该基金利用灵活的投资战略赚取收益，常常存在于富有的投资者持有的证券投资组合中。因为通常来说，只有“合格的购买者”“合格的客户”或者“可信的投资者”——最起码资本净值得超过100万美元，才可以在《美国联邦证券法》的允许下投资对冲基金。当然，即使你不富有，你仍然很可能对对冲基金感兴趣。比方说，如果你是得克萨斯州的公务员或者退休职工，你的养老金中有大约5%可以用来投资对冲基金。[注]

对冲基金经理的工作像多数投资者的工作一样，是预测下一步会发生什么。他得根据对冲基金的目标买进卖出投资项目。而且，至少有一项研究表明，即便女性对冲基金经理要比男性更出色，对冲基金经理通常都是男性。[注]

问题是预测的准确与否取决于其所依据数据的质量。

在我们采访到Eagle’s View资产管理（基金内的对冲基金）创始人及总裁尼尔·伯杰时，他说：“我认为，在华尔街，可以根据10年间的收益得出结论这样的想法荒谬至极。”[注]基金内的对冲基金是指投资不同的对冲基金，就像投给不同股票的公共基金一样。“哪怕向前推100年，都不足以从统计学角度得出一条正态分布曲线，”伯杰补充道，“我们还未掌握充分的历史数据来预测股票市场的走向。”

当然，在谈到对冲基金和预测时，有一点很有趣，即对冲基金（可能因为管理费及其他往往与之相关的费用）并不总能比股票市场提供更好的收益。正如《彭博商业资讯》的一篇头条文章所提出的：“对冲基金的收益连续5年比股票收益低7.4个百分点。”[注]这可能是典型的筛选数据，因为看其他时期的数据时会得出截然不同的结论——包括一篇《华尔街日报》上的文章也称：“15年来，（对冲基金的）收益超过了整个股票市场的收益。”[注]公平地讲，正如伯杰及其他投资家所指出的，在正常收益上超出标准普尔500指数（一份大约有500只且不断变化的股票名目）可能不是所有对冲基金的目标。而最终目标往往是提供最好的风险调整收益——要采取调整风险的措施来获取收益。尽管有的时候预测会出现偏差。在一个经典的案例中，对冲基金长期资本管理公司（LTCM）不到一年的时间“从47亿美元中损失了44亿美元”，造成该结果，部分归因于利差关系同预期不相符。[注]

暂且不论它们的表现，对冲基金有时会因为对冲基金经理所获得的薪资受到指责。《机构投资者的阿尔法》（*Institutional Investor’s Alpha*）发布了年度“富豪榜”，评估了高级经理的收入。甚至在强调了近10年来对冲基金经理收入最少的2015年的榜单上，25名上榜人员“加起来共赚了微不足道的116.2亿美元”。那是他们之前收入的一半，你可能会为他们感到难过（尽管平均4亿美元的薪资就已经很难赚了）。[注]

显然，预测正确的话，可以获得丰厚的收入。

当然了，你不必成为一名受惠于预测的对冲基金经理。比如，正如经济学家J.J.普雷斯科特（J.J.Prescott）在一次访谈中所解释的那样，谈判往往非常倚重预测。普雷斯科特指出："在谈判中，'BATNA'（达成谈判协议的最佳选择方案）是谈判人员必须了解的头等大事。如果对方有好的外部选择（像诉诸法庭），那么以令人满意的方式解决问题就变得困难了。因此在和解谈判中，你不仅需要预测自己的最佳备选方案，还要判断对手的最佳备选方案是什么。"

思考一下，一个决定将会如何对你、你的同事、你的对手、你的竞争者产生影响。这样的思考将会帮助你在职场中占据优势地位。

---

1. "Performance," Employees Retirement System of Texas website, accessed July 23, 2015, http://www.ers.state.tx.us/About_ERS/Investments/Performance/.
2. "Women in Alternative Investments: Building Momentum in 2013 and Beyond," Rothstein Kass Institute Second Annual Survey, December 2012, https://nyhfr.org/documents/FG/hfrt/edu/70046_RK_WomeninAlternativeInvest mentsF.pdf.
3. Full disclosure: Mike's brother-in-law works for Eagle's View.
4. Kelly Brit, "Hedge Funds Trail Stocks for Fifth Year with 7.4% Return,"Bloomberg Business website, January 8, 2014, http://www.bloomberg.com/news/articles/2014-01-08/hedge- funds- trail- stocks- for- fifth- year- with-7-4-return.
5. Rob Copeland and Gregory Zuckerman, "How Individual Investors Can Invest Like a Hedge Fund," Wall Street Journal website, August 3, 2014, http://www.wsj.com/articles/ how- individual- investors- can- invest- like-a- hedge- fund- 1407106285.
6. Stephanie Yang, "The Epic Story of How a 'Genius' Hedge Fund Almost Caused a Global Financial Meltdown," July 10, 2014, Business Insider website, http://www.businessinsider.com/the- fall-of- long- term- capital- management- 2014-7#ixzz3kjpsJKfl.
7. Stephen Taub, "The 2015 Rich List: The Highest Earning Hedge Fund Managers of the Past Year," Institutional Investor's Alpha website, May 5, 2015, http://www.institutionalinvestorsalpha.com/Article/3450284/ The- 2015- Rich- List- The- Highest- Earning- Hedge- Fund- Managers-of- the- Past- Year.html.

你知道什么？

抛硬币的结果是确定的。所以你可以尽情预测，但是你知道结果只能三选一：正面朝上、反面朝上或者（如果你较真的话）立起来。

许多预测会仅仅因为其实施方式，导致结果缩小到一个已知的范围。比如，如果你试图预测以下情况，那么你的选择有限：

•谁会赢得世界职业棒球大赛。

•在新学年开始的时候预测期末微积分的成绩。

•猜拳时，你的对手会出石头、出布，还是出剪刀。

从另一方面来看，在有些预测中，可能出现的结果将会多很多，甚至可能有无数种可能。比如，如果你试图预测以下情况，那么可能出现的结果将会多得多：

•50年后，谁会成为美国总统。

•无线电台播放的下一首歌会是什么。

•10年后你的职称（还需考虑到未来的一些工作在当前可能闻所未闻）。

如果你能够限定结果的数量，你的预测可能会发生很大变化，也许还可以降低预测误差。每年职业棒球大联盟球队无缘季后赛时，我们就会遇到这种情况。每淘汰一支球队，留下的每支球队赢得世界职业棒球大赛的可能性就大大增加。正如预测的时间范围越短，预测的结果越精准一样（其他条件相同），在有些情况下，通过限制潜在结果的数量，

你可能会得出更准确的预测。然而，即便有的时候结果只有两种可能，老练的民调分析员都可能估计错误。

民调为什么会出错？

多数人认为，托马斯·杜威（Thomas Dewey）本可以成为美国第34任总统。

1948年，重要的民调预测纽约州长杜威会在总统选举中赢得压倒性胜利，但选民们没有将杜威送进白宫，而是以近5个百分点的优势选了哈里·杜鲁门做总统。

这件里程碑式的事件导致了20世纪让人最难以忘怀的媒体乌龙之一，当时《芝加哥论坛报》印刷了近15万份带着错误标题“杜威击败了杜鲁门”的报纸。[注]

那么这是如何发生的呢？民调为什么错了？

一个因素是，据《洛杉矶时报》报道，当时盖洛普民调组织联合主席小乔治·盖洛普（George Gallup Jr.）称，民调者“提早几周”就停止调研了。[注]民调者认为在选举前的最后几周民众的意见不会有太大的变化了——然而实际上有变化，因为杜鲁门的支持者集合了工人们的选票。与此同时，民调者彭斯·W.“巴德”·罗珀指出，自负的杜威支持者们“当时都打高尔夫球去了”。[注]

另一个因素是，民调模型建立在以往选举的基础之上，1948年竞选之后，选举在很多方面都发生了变化。那时，1932年后的每次总统选举——16年的时间——富兰克林·罗斯福都参选并赢得选举。在每次选举中，罗斯福都会击败一个主要对手。直到1948年——这场没有罗斯福参与的竞选有四位主要的竞选者（杜威、杜鲁门、斯特罗姆·瑟蒙德、亨利·华莱士）。[注]

这次，报社做了确定性预测，提前声明杜威将会赢得选举。但是从民调得来的数据显然没有将选民最新的情绪考虑在内，而且当时的统计模型似乎是围绕着二人竞选而创建的。我们就是这样得出《时代》杂志所说的“公认的美国政治史上最具颠覆性的事件”的。注

---

1. Tim Jones, “Dewey Defeats Truman,” Chicago Tribune, accessed July 16, 2015,http://www.chicagotribune.com/news/nationworld/politics/ chi- chicagodays- deweydefeats- story- story.html. Perhaps you’ve seen the famous photo of Harry Truman holding up a copy of the paper, a photo taken two days after the election— a nearly unimaginable span of time in today’s digital world. That said, the error was also due, in part, to a printers’ strike, which meant that inexperienced typesetters were working at the paper on election night. In fact, on the same front page with the “Dewey Defeats Truman” headline, you can see five lines of upside- down type in the far- right column: Todd Andrlik, “Dewey Defeats Truman: The Rarely Told Story of Chicago Tribune’s Most Famous Issue,” Huffington Pos website, updated May 25, 2011, http://www.huffingto post.com/ todd- andrlik/ dewey- defeats- truman- the_b_119351.html.
2. Will Lester, “‘Dewey Defeats Truman’ Disaster Haunts Pollsters,” Los Angeles Times website, November 1, 1998, http://articles.latimes.com/1998/nov/01/news/mn-38174.
3. Lester, “‘Dewey Defeats Truman’ Disaster.”
4. Thurmond and Wallace each pulled in 2.4 percent of the popular vote nationwide, although — given the nature of the U.S. Electoral College system—Thurmond carried four Southern states, while Wallace carried none. The 1948 race had all sorts of interesting twists and turns. Thurmond got one Electoral College vote from Tennessee, as faithless elector Preston Parks decided not to vote for Truman, who had carried the state. (Faithless electors— those who don’t follow their party’s wishes— are an example of a variable that can disrupt predictions.) In addition, due to efforts by Dixiecrats, Truman was left off the ballot in Alabama, despite the fact that he was the sitting president and the official Democratic Party nominee.
5. Ben Cosgrove, “Behind the Picture: ‘Dewey Defeats Truman’ and the Politics of Memory,” Time website, May 4, 2014, http://time.com/3879744/ deweydefeats- truman- the- story- behind- a- classic- political- photo/.

偶然与概率

一如往常，留意数据的使用方式很重要。比如，假如今天晚上百分之百可以吃到比萨与下周百分之百可以吃到比萨，从统计学上讲，二者截然不同。前者，你知道你有一晚会吃比萨，而后者，你也许每晚都吃比萨——或者7晚当中，你只有一晚吃了比萨。

尽管有时人们会将概率（probability）和比率（odds）二者互换使用，但是二者不一样，人们常常误用这两个词。概率是某一结果发生的可能性。某事发生的比率，从统计学上讲，是指有利结果与不利结果二者的比例。

拿一副52张的标准扑克牌来说明吧。

你抽中人头牌（J牌、Q牌、K牌）的概率为23%，一副牌中有12张人头牌，12除以52得出23%。

抽中人头牌的比率为12∶40（一副牌中有12张人头牌和40张非人头牌）。

概率和比率二者是有联系的——你可以从其中的任何一个推算中另一个，但它们是两回事。

在我们谈论未来事件时，你可能也会听到“风险”这个词。风险一词的含义因人而异，故需要谨慎对待——若你从事的是投资或者保险业，就尤其需要注意了。不过一般来看，风险会把事件的概率及后果均考虑在内。

比如，迈克知道怎样玩杂耍。当他表演扔苹果时，漏接一个苹果，

苹果砸到他脚上的概率为1%；当他表演扔刀时，漏接一把刀，刀砸到他脚上的概率也为1%。概率相同，但是风险却不一样。（《纽约时报杂志》中有一篇题为“在大自然的赌场中”的文章，是写灾难风险的，文章写得十分有趣，让人读得惊心动魄。你可以读一下，了解低估飓风带来的危害是如何让美国保险业濒临破产的。[注]）

一件事情发生的概率较低并不意味着它不会发生。《飞翔》杂志的编辑在《大西洋报》的一篇文章中称，NASA对某些部件定下的标准为“发生故障的概率为$1/10^9$，或者说是10^{-9}。NASA称这样的部件‘不可能发生故障’”。[注]然而它并不是不可能。实际上，如果数据统计无误，概率为$1/10^9$的事情将会在每10亿件事情发生后出现一次。正如伯杰所说的，赌场是不会让比尔·盖茨一次下50亿美元的注来赌的。“如果他们是世界上最不幸的人，他们会因此而破产。”

尤其在你着手去看大数据集时，你可能期待看到这些低概率事件。正如《华尔街日报》在一篇关于彩票抽奖中巧合事件的文章中所述：“每周全世界会有数百万人从成百上千张彩票中选取数字，巧合是肯定会发生的。”[注]

想想黑天鹅。几百年前，人们认为不可能存在黑天鹅，仅仅是因为他们之前没有见到过任何黑天鹅存在的证据。但是看不到黑天鹅并不意味着黑天鹅不存在，只是我们还没见到而已。现在，“黑天鹅”事件就是指极其不可能发生，但一旦发生会带来巨大影响的事件。该词是由在不确定性这一话题上著述颇丰的纳西姆·尼古拉斯·塔勒布推广开来的。仅仅因为它还未发生不意味着它不能或者不会发生。黑天鹅确实存在。[注]

---

1. Michael Lewis, “In Nature’s Casino,” New York Times Magazine, August 26,2007, http://www.nytimes.com/2007/08/26/magazine/26neworleans-t.html?pagewanted=all&_r=0.

2. James Fallows, “When a 1-in-a-Billion Chance of Accident May Not Seem‘Safe Enough,’ ”Atlantic website, March 28, 2014,

http://www.theatlantic.com/technology/archive/2014/03/when-a-1-in-a- billion- chance-of-accident- may- not- seem- safe- enough/359780/.

3. Carl Bialik, “Odds Are, Stunning Coincidences Can Be Expected,” Wall Street Journal website, updated September 24, 2009, http://www.wsj.com/arti cles/SB125366023562432131, accessed August 2, 2015.

4. Taleb cites the rise of the Internet and the events of September 11, 2001, as examples of events with black swan characteristics in his book The Black Swan:The Impact of the Highly Improbable, 2nd ed., with a new section: “On Robustness and Fragility” (Incerto), Random House (2010).

心理因素影响预测

到现在为止，我们主要关注的是数据概念，但确实存在许多影响预测的心理因素。比如：

•自负——《财经日报》（*Journal of Finance*）的一篇文章发现，高成交量（有时可以归结为自负）会导致糟糕的财务状况；在股票市场回升近18个百分点时，交易最多的人获利不到12个百分点。(注)

•恐惧——根据《周刊》杂志统计，被鲨鱼咬死的概率为1/（3.7×10^{6}）。(注)但当你和亲友在沙滩上，你脑海中不停地回响着《大白鲨》中的音乐时，你很难进行有逻辑的思考以及精准的预测。

•评定习性效应——正如《哈佛商业评论》中的一篇文章所述，人们在对某人做出评判的时候，影响评判结果的实际上是评判者而非被评判者。文章发现，通常“我对你的评价中，有61%的内容是对我自身的反映”。(注)

预测未来并非易事。也许这就是电话灵媒收费昂贵的原因了。《宋飞正传》试播集的讨论组说它“竞争力不强”，然而这档节目继续播出，并成为全天收视率最高的电视剧之一。(注)数据发挥着关键作用。但是自负、恐惧以及古而有之、简单直接的人类本性则是影响预测的几个非数据因素。留心它们，会让你把工作做得越来越好。

---

1. Brad M. Barber and Terrance Odean, “Trading Is Hazardous to Your Wealth:The Common Stock Investment Performance of Individual Investors,” Journal of Finance LV (April 2, 2000), http://citeseerx.ist.psu.edu/viewdoc/download?d oi=10.1.1.139.1931&rep=rep1&type=pdf.
2. Harold Maass, “The Odds Are 11 Million to 1 That You’ll Die in a Plane Crash,” The Week website, July 8, 2013, http://theweek.com/articles/462449/odds- are-11-million-1- that-

youll- die- plane- crash.

3. Marcus Buckingham, “Most HR Data Is Bad Data,” Harvard Business Review,February 9, 2015, https://hbr.org/2015/02/most-hr-data-is- bad- data.

4. “The Seinfeld Chronicles: An Obsessive- Compulsive Dissection of All 169 Episodes,” Entertainment Weekly website, May 4, 1998, http://www.ew.com/article/1998/05/04/ seinfeld-chronicles.

如何成为一个聪明的预测者

从你睁开眼在智能手机上看到的天气预报到晚间新闻上最新的大选民调，预测无处不在。为了弄明白预测，你可以从以下5件事情做起。

1.知道预测未来取决于对过去（或者模型）的了解。如果以往数据或者模型出现了数据问题——抽样误差、遗漏变量、误算平均值、结构变化等——这些问题将在预测中一一暴露。

2.存在不同类型的预测——尤其是，确定性预测和可能性预测。当你观察一项预测时，弄清楚它的类型。是在说你明天的高尔夫锦标赛期间会下雨吗（确定性）？还是说比赛有40%的可能性会因为下雨而取消（可能性）？

3.了解术语。预测和预知常常是同义词，但并不总是如此。当人们使用“可能性”“概率”“风险”“比率”等词时，你得明白他们在谈论什么。对于经济学家和地震学家或者预言家和保险公司高管来讲，相同的词可能意味着截然不同的事物。

4.知道预测的准确性可能会随着时间变化。在第七局时预测棒球赛的最终得分比一开始就预测得出的结果更精准，因为你有更多可用数据。当研发出搜集数据、创建数据模型的新技术、新方法时，预测也会变得更加精准。只需想想基因药物所有的最新进展，这些新进展赋予了我们对人类长期健康做出预测的能力。也就是说，让你改变预测的机会越小，你所做的预测就会越精确。

5.允许某种程度的不确定性的存在。正如对冲基金人尼尔·伯杰所说的，“我们不得不在一个永远不会100%确定或者100%舒适的世界中生活”。让人们甘心认为我们生活在一个不可预知的世界里困难重重。很

多人想不明白，认为每天日子照常过。但是我们无法降低每一次可能的风险。我们必须尽我们所能做出最佳判断。

纽约洋基队的著名球员尤吉·贝拉曾说过这样一句经典的话：“做预测本就不易，预测未来更难。”

第九章 拨开数据的迷雾：总结

想象一下，你去动物园然后看到只狮子，远处笼子里有头大象，另外的展区有长颈鹿，等等。每一种动物都在自己的隔离区域，每个区域有一个漂亮的小标签，告诉你更多有关动物的信息。

这就是我们在本书中采用的方式，每次解释一种数据概念，每种概念单独成章。

现在，是时候回顾一下了。

在现实世界中，数据问题几乎不会孤立存在。你会碰到多个数据问题同时出现的情况，比如，仅仅因为你鉴别出了抽样误差并不意味着筛选数据、关联性、平均数、预测等问题不存在，抑或只是存在更多针对这件事的抽样问题。有的案例可能没有数据问题，有的则有很多，你需要睁大双眼看清所有问题。

这就是本章涉及的全部内容。

在此，你会发现一些我们最喜欢的研究和故事，这些研究和故事展示了不同的数据概念是怎样影响分配、消费、诠释数据的。

那我们开始吧。

不要轻信头条新闻中的数据

《大西洋报》上的一篇文章提出了这样一个问题，“为什么前高中运动员收入更高？”㊟据该文章称，与高中时代不是运动员的人相比，高中时代是运动员的人最终“职位更高”，薪资比当年可怜巴巴的长号手及年鉴俱乐部主席高“5%~15%”。

好吧，好像高中时代吹长号的那些孩子还不够惨似的，毕业后还要继续不如别人。该文章吸引了“小数据夫人”（约翰的妻子）的注意，因为这类问题会使成熟的数据接收者不满，在她的要求下，我们调查了该研究及基本数据。

文章是基于康奈尔及南伊利诺伊大学研究员对两组独特的履历资料集的研究而撰写的。㊟从心理学角度看，履历资料就是一种自陈式数据（例如传记数据）。有两项研究——第一项研究抽选了66名成年人，让他们参与一项关于领导力特质及以往课外活动经验的调查。这部分调查的目的是收集人们对参与体育活动，以及课外活动是如何与诸如自信、领导力、自尊等品质产生联系的看法。

从这66人组成的样本中，文章作者得出了以下结论：“人们倾向于认为前学生运动员会表现出更强大的领导能力及益于组织的个人品质，但同时不会心怀尊重地为他人着想。”对于这项研究，我们不想花费太多时间，只想说这类关乎个人观点的调查往往会在不同领域引起争议（经济学家往往表示严重质疑，而产业组织心理学家则经常对这类调查数据加以利用）。然而，从统计学角度看，之所以得出了这样的结论，原因在于将所有体育活动和非体育活动泾渭分明地区别对待。

根据上面的说法，我们来关注一下第二项研究，该研究就伊利诺伊

大学对2000名老兵的调查进行了数据分析。这项研究中，样本包含了历经二战的931名老兵的信息，在2000年调查完成时，他们最年轻的有71岁，最年长的有93岁。研究报告撰写者得出的主要结论是，自孩童时就参与体育运动，其积极影响会“持续55年以上”。根据撰写者的描述，“参与青少年竞技体育与个人长远成功与亲社会性的几项指标之间呈正相关关系”。

下面列出几项针对上面研究的调查结果。

首先，研究一和研究二都没有观察或者衡量高中时代的体育运动对薪资造成的实际影响。这类数据并不存在。而这两项研究观察的是自55年前大学体育活动参与情况的自陈式数据与领导力指标、贸易工作、高级管理工作是如何相关的。对于考量一个人是否找到更好的工作，这样的研究方法有失偏颇。

其次，这两项研究的前提是55年前参与的体育活动和之后找到什么样的工作相互关联。而研究过程中，可以解释的变量多少是受限制的；实际上，这两项研究可以掌控的与结果有关的其他解释只有年龄及家乡的规模。这就引出了遗漏变量的问题：倘使老兵中71~93岁男性老兵这一子集的高中体育参与情况与其他因素有关联——如在《退伍军人法》下享受的福利、受教育程度、二战参与度、个人能力等，结果又会怎样呢？在许多对工作表现的研究中，这都是一个经典议题。

最后，老兵研究报告的撰写者还提出，他们的研究结果暗指关联性而非因果性。这项研究有很多双向分析——领导力与体育运动、自信与自尊等。接下来是一个可能的解释：倘若有时间进行体育运动的孩子都是那些家境富裕的（比如放学后不必去打工）会怎样呢？小时候家境富裕可以对其之后的成功做出解释。关键是，这个问题纷繁复杂，寄希望于通过这些特殊指标来捕捉真正的因果关系几乎不可能。

我们还没有把生物特征数据考虑进去——尽管这也是一项有趣的研

究。但上文这个例子，又一次说明从头条新闻中获取的数据并不能完全传达对基本研究的微妙阐释。

---

1. Joe Pinsker, “Why Do Former High- School Athletes Make More Money?,” Atlantic, May 28, 2015, http://www.theatlantic.com/business/archive/2015/05 /why-do- former- high- school- athletes- make- more- money- and- get- better-jobs/394283/.
2. Kevin M. Kniffin, Brian Wansink, and Mitsuru Shimizu, “Sports at Work.Anticipated and Persistent Correlates of Participation in High School Athletics,”Journal of Leadership & Organizational Studies 22, no. 2 (May 2015): 217– 230,doi: 10.1177/1548051814538099.

如何看待工作满意度调查？

“花多少钱才能让律师开心？”

这不是开玩笑。它是法律360（Law360）网站上一篇关于律师满意度调查的文章的标题。

说真的，要花多少钱呢？答案是什么？每年50万美元。

据法律360网的研究，收入超50万美元的律师在工作中更开心，跳槽的可能性也较小。正如文章所述，“尽管幸福与薪酬高低没有直接联系，但50万美元似乎成了一个门槛，超过它，律师们会认为他们对工作的方方面面‘满意’或者‘非常满意’”。[注]

法律360网还在一连串与调查相关的信息中发现：

•职业律师最开心。

•同非股权投资公司合作的律师“最悲催”。[注]

•500人以上大公司的律师比小公司的律师开心。

这些发现可以成为头条，而且必定会在从华盛顿到达拉斯的律师事务所玻璃会议室内引起一阵骚动。

但是在你为此争得面红耳赤之前，我们不妨仔细看看这个数据的真正含义。

美国有130多万名律师。[注]调查根据的是其中300名律师的回应。不过我们更关注的不是样本大小而是样本构成。因为，根据法律360网的

文章来看，结论似乎基于回应调查的、自我归类的律师。在你让一组人选择参与到研究中时，有可能出现选择偏差。结论可能偏向于选择参与的那部分人。做出回答的律师与那些因为过于忙碌而无法回答，或者由于某种原因选择不回答的律师有区别吗，这些区别与调查结果有关吗？比如，有可能出现忙并快乐着的同非股权投资公司合作的律师没有时间回应调查的情况吗？当然可能。

由于这是自陈式数据，律师在调查中故作开心、抬高薪资，提供各种各样错误信息的情况也可能发生。我们有证据证明他们这么做了吗？没有。但是在这类调查中，并没有适当的手段来保证不出现误导人的（或者完全错误的）回应。自陈式数据本身未必是问题，问题是我们可能因为自陈式数据受到人为操纵而得到错误的答案。

我们来看看法律360网所说的比业余律师幸福的职业律师。不过职业律师——至少回应该调查的那些——没有业余律师赚得多。鉴于薪资被认为是关乎幸福的，也许还有其他变量在此发挥作用？抑或研究劳动法的人只是自然而然地更加积极乐观。遗漏变量是什么呢？我们不知道——这是个问题。

我们也不知道金钱是否真的“令律师开心”，正如前文提到的标题所述，在我们还未看出证明这不只是关联性的证据时，就暗示了一种因果性。实际上，在另一个故事中，法律360网指出“幸福不完全是由金钱主导的”，这就增加了金钱之外推动幸福的因素存在的可能性。[注]

还有一个问题——缺少上下文。如果你读完所有文章，你可能了解到其中一组律师的满意度比另一组高17%，或者留在公司的可能性高32%等。但是法律360网似乎没有就其理解结论含义的方法给出足够多的细节。律师是否被要求用1~5（也可能用1~100或者其他范围）为自己的满意度打分？还是为他们提供了诸如“不满意”“满意”“非常满意”的选项？掌握更多调查方法的信息可以更深入地了解调查结果的真正含义，

还可能发现调查方法中的任何偏差。

发布这类结论的肯定不只法律360网一个。我们已经在计算机、建筑、工程及许多其他行业中见过类似的调查了。我们不是说我们期望这类文章要写得多严谨，但在你阅读邮件、浏览智能手机上的新提醒、接收工作及家庭中环绕你的数据时，希望可以思考一下我们提出的这些问题。

---

1. Aebra Coe, “How Much Money Does It Take to Make A Lawyer Happy?”Law360 website, August 17, 2015, http://www.law360.com/articles/691712/how- much- money- does-it-take-to-make-a- lawyer- happy.
2. Andrew Strickler, “Nonequity Partners the Most Miserable Attys in Your Firm,” Law360 website, August 17, 2015, http://www.law360.com/articles/691856/ nonequity- partners- the-most- miserable- attys-in- your- firm.
3. “ABA National Lawyer Population Survey. Historical Trend in Total National Lawyer Population 1878– 2015,” American Bar Association, accessed August 17, 2015, http://www.americanbar.org/content/dam/aba/administrative/market_research/ total- national-lawyer- population- 1878- 2015.authcheckdam.pdf.
4. Jacob Batchelor, “Bigger Is Better When It Comes to Making Lawyers Happy,”Law360 website, August 17, 2015, http://www.law360.com/articles/691835/bigger-is- better- when-it-comes-to- making- lawyers- happy.

出生月份与健康有关系吗？

“占星术可能是假的，但你出生的月份似乎真的与你的健康有关。”《华盛顿邮报》如是说。注

此话当真？

我们同意占星术有些方面不真实，不过你出生的时间真会影响你的健康？

你首先要做的事情就是思考该研究是不是被媒体误解了。之前我们确实遇到过——有一篇科学论文探究了二者的关联性，随后被媒体添油加醋一番，成了二者暗含因果性。然而，在这个例子中，《华盛顿邮报》的报道如往常一样似乎是真实的。文章仔细地解释了你出生的时间与你今后可能罹患的“疾病之间存在联系”。而当我们去看《美国医药信息学协会期刊》（*Journal of the American Medical Informatics Association*）上的原始研究时，其证实了我们在《华盛顿邮报》上看到的内容。研究指出：“一个人的出生月份对今后所患疾病有着重要的影响。”注

那好，接下来的一步是试着弄明白出生月份与疾病二者存在联系的原因。跟往常一样，我们从调查样本开始。尽管研究员查看了175万人的记录（样本容量似乎很充足），但是他们只抽取了在曼哈顿的纽约长老会医院及哥伦比亚医学中心接受治疗的病人。自不必说，曼哈顿代表不了整个纽约城，更不用说整个美国了——或者往大里说，全球人口了。比方说，仅有1%的病人被认定是亚洲人（尽管在占了样本容量近50%的“其他”及“未识别”分类中可能会有更多人）。

那么，样本中的内容能够解释出生月份与疾病的关系，这可能吗？

倘若在元旦前夜，时代广场上的五彩纸屑引发了与出生月份相关的健康问题——哮喘，会怎样呢？这正是你需要考虑的存在遗漏变量的情况。换句话说，出生月份是导致一些人疾病高发的原因吗？或者说出生月份只是与疾病具备关联性，而真正引发疾病的是其他因素？例如，正如笔者所言，季节性因素（过敏、晒太阳等）也可能会发挥作用。

作为一个明智的数据接收者，你还考虑到了什么？下面是一个有趣的观察——在研究中，研究员发现了同出生月份“息息相关”的55种疾病。单独来看，数目巨大。但是研究员调查了1688种不同疾病的数据，这意味着，在他们研究的疾病当中，仅有约3%的疾病与月份有联系。有时，鉴于显著性往往衡量着某种关系的概率，你研究的数据越多，你越有可能发现具有显著性差异的结果。出生月份和疾病之间的关系是不是随机可能性造成的呢？有的P值小于0.001，而有的则非常高。这种结果可能是我们之前讨论过的多重比较所导致的。

那么现在我们还有个问题：我们该如何处理这些数据？我们来看一下影响的尺度（大小）。在有些情况下，出生月份可让患病风险升高约5%，或让寿命缩短或者延长四五个月。[注]这足够让人忧心了吧？

最后，我们来思考一下这个数据会对你的生活产生怎样的影响。研究者提到的55种疾病都会对人们的生活产生不同的影响。比如哮喘，可以通过药物或者改变生活方式得到治疗，而一些心血管疾病（像充血性心力衰竭）处理起来要更加周密细心。除了个人的健康问题，你会如何运用这些发现来决定何时生孩子？如果你在约会，你会问对方的生日是什么时候，以过滤掉潜在的不健康的追求者吗？像很多研究一样，这项研究也提出了诸多问题，不仅与数据的统计学影响有关，还涉及对我们生活造成的经济影响。

---

1. Ana Swanson, “Scientists Have Discovered How the Month You’re Born Matters for Your Health,” Washington Post website, June 15, 2015, http://www.washingtonpost.com/news/wonkblog/wp/2015/06/15/ what- your- birth- month-

means- for- your- risk-of-disease/.

2. Mary Regina Boland, Zachary Shahn, David Madigan, George Hripcsak, and Nicholas P. Tatonetti, “Birth Month Affects Lifetime Disease Risk: A PhenomeWide Method,” Journal of the American Medical Informatics Association (June 3,2015), doi: http://dx.doi.org/10.1093/jamia/ocv046.

3. The reasons vary based on the medical condition. For example, the researchers(citing other studies) noted that “individuals born in seasons with more abundant home dust mites had a 40% increased risk of developing asthma complicated by dust mite allergies,” given that “sensitization to allergens during infancy increases lifetime risk of developing allergies.”

如何解读关于问题学生的数据？

哪怕在中学，孩子们都知道“耍酷的学生”都是什么德行。他们会看少儿不宜的电影、会从当地便利店偷零食，还会躲在学校后面抽烟（或者更带劲儿的东西）。

但是这些惹是生非的青少年终究会改掉这些陋习的，对吧？难道中学时期的小混混就没有希望了吗？通过你13岁时的所作所为真的能预测到23岁的你何去何从吗？

如果你的孩子是少年犯，你可能就不想看接下来的内容了。因为研究者发现，孩子青春期早期假装成熟的行为（孩子努力表现得比实际年龄大）“预示着长期的亲密关系障碍、严重的酒精及药物使用问题，以及犯罪行为严重程度的飙升”。[注]太吓人了。

该研究——第一段首先借莎士比亚、詹姆斯·迪恩、蒂娜·菲之名来抬高自己——发表于《儿童发展》，随后还被多个媒体报道，鉴于所讨论的话题颇具争议性，所以这样的反响不足为奇。在新闻发布会上，主持该项研究的教授试图解释早期行为与后期结果之间的联系。从理论上说，随着时间推移，耍酷的孩子“需要用越来越极端的行为来表现自己很酷”。[注]

此时，假如你是家长，你可能读到这项研究结果，并想把孩子锁在房间里直到他们30岁。但是作为一名数据接收者，你可能就会从其他几个角度来看待这项研究。

比如，你可能想知道这些青少年是如何从全国人中挑选出来的——抽样。研究者研究了公立学校中来自美国东南部城区及郊区的孩子，他们的家人说他们对这项研究感兴趣。同样的结论适用于那些在家学习或

者在私立学校读书的孩子吗？东南部以外地区的孩子们当中也能发现相同的行为模式吗？那些家人没兴趣参与这项研究的孩子们，又是如何？在这种情况下，我们没有可以明确证明结论会如何不同的数据。

在研究涉及的6个地区中，研究人员谈到了求数据平均值的方法。比如，研究者选取了多年以来酒精摄入的数据，通过取平均值来得出一个总体的药物使用得分。是的，在一些案例中，数据是自陈式的。

研究还突出了另外一个值得注意的问题——把从一个样本预测出的结果应用到一个不同的样本集或数据总体中。媒体报道这些类型的研究时，有时会发生这类情况。比如，一家媒体（出于保护目的，暂不提名了）刊登了一篇讲述中学时代的行为如何能够影响“你们的未来”的文章。这篇文章耍了点小聪明，文章并没有说184名青少年的行为如何预示他们的未来，而是以你们为对象——是说给读者听的。

当媒体报道或者原始研究开始将结果归因于样本集之外的人或者其他主体时，那就涉及了外延有效性——这是我们衡量结论是否适用于样本之外的一种方法。有的时候适用，有时候不适用——但这是你需要注意的事情，特别是在读新闻的时候。

---

1. J. P. Allen, M. M. Schad, B. Oudekerk, and J. Chango, “What Ever Happened to the ‘Cool’ Kids? Long- Term Sequelae of Early Adolescent Pseudomature Behavior,” Child Development 85, no. 5 (September/October 2014): 1866– 1880.

2. “New Study Sheds Light on What Happens to ‘Cool’ Kids,” Eureka Alert!, Public Release, June 12, 2014, http://www.eurekalert.org/pub_releases/2014-06/sfri- nss060514.php.

如何使用房价评估网站的数据？

如果你想买卖房子，应该对Zillow网有所耳闻。凭借其容纳美国超过1.1亿家房子的数据库，Zillow网几乎能够让你看到美国任意一套房子的售价。(注)

多多少少可以这么说吧。

这就是它的运作模式：在你寻找房子时，你会看到“Z评估”（Zestimate），它被冠以Zillow好听的品牌名，主要用于房产市场价值评估。

Zillow网运用了依托于公开数据及用户提交数据的专有公式得出Z评估的结果。根据其网站的说法，该公司掌握了有关房产的各种数据，包括：

•建筑面积。

•位置。

•房间及卫生间数目。

•房产税数据。

•同区域内其他房产的可比销售情况。

实际上，你现在就可以在Zillow网上查看自家房屋，看看Zillow网认为你家房屋值多少钱——即便房子是几年前买的都行。

如果没有足够的数据来得出Z评估，Zillow网有时候就干脆不提供

评估了。例如，在我们查看Zillow网的数据时，佛蒙特州没有一栋房子有Z评估资料。(注)完全没有相关信息。如果你住在新泽西，你就幸运了——Zillow网为“花园之州”99.4%的房屋做了Z评估。

即便Zillow网指出Z评估并非对价格的评估，许多人在买卖房屋的时候仍然依赖它。《洛杉矶时报》上的一篇文章指出“如果在售房屋的Z评估为35万美元，买家可能会对卖家价目表上的42.5万美元产生怀疑”。(注)类似地，卖家可能会问经纪人为什么他们房屋的价目表会远远低于Z评估。(注)

“每周我们都会遇到因为从Zillow网上看到某个参考价格，就死死咬住不放的顾客。”

一家不动产经纪公司的工作人员在为《华盛顿邮报》撰文时写道，他还在文中称Zillow网的估价“极不准确且反复无常”。(注)

在《华盛顿邮报》的一篇辩驳文中，Zillow网首席经济学家说，人们不可能100%确定地预测出一座房屋的销售价格，同时还承认一项华盛顿哥伦比亚特区的研究中所说的，不动产经纪人的表现胜过Zillow网。他问道：“一个消息灵通的人比电脑更擅长估算私人房屋价格吗？”“当然，答案是肯定的。但两者所估算出来的价格比你想象的要接近。”(注)

所以你该如何解读以上所有这些信息呢？

随着你开始深入挖掘数据，你将会发现，Zillow网为每一个地区设定了一个名为平均错误率的东西。平均错误率是一个百分数，它告诉你在这个区域有一半的Z评估价格比算上错误率更接近最终销售价格，另一半则相差较远。

比如，当我们在看洛杉矶地区的数据是，平均错误率为7%。这意

味着一半的房屋售价与Z评估的估价差距小于7%（记住，这里的7%既可以表示更贵，也可以表示更便宜——所以一座售价80万美元的房子价格会在74.4万美元和85.6美元之间波动）。当然，平均错误率也意味着一半在售房屋会以这个范围之外的价格售出——要么小于Z评估的7%，要么大于7%。[注]

在有些情况下，Zillow也提供它所谓的“价值区间”，价值区间表明房屋的预估最高价和最低价。正如其网站所解释的，这个区间实际上一个70%的置信区间，意味着在70%抽样中，房屋价值应该在该区间之内。

换句话说，Z评估应该将其本身的平均错误率考虑进去。值得称赞的是，Zillow网公开了平均错误率并对其加以解释。它甚至通过四星评价系统简化了全部操作，来告诉你Z评估有多么准确（如你所预料到的，星级似乎与平均错误率一致）。[注]问题是，许多人仅仅关注Z评估而忽略了其背景，正如许多人看民调时不考虑误差范围一样——他们只看谁将会“胜出”。然而通过忽略附加数据——不管是平均错误率、误差范围还是置信区间，你没有从大局着眼。假设你住在华盛顿哥伦比亚特区，在Zillow网罗列的顶级都市圈中，该地区的Z评估精确度最高。[注]如果你试图卖掉一座Z评估为50万美元的房屋，尽管你的房屋估价的平均错误率在5%以内（记住，只有一半的房子如此），那意味着房价可能会比预估值多或者少2.5万美元（50万美元的5%）。那就是5万美元的区间。

Z评估也很有趣，因为有的人会将它们当预测来用，试图判断在未来房屋会卖出什么样的价钱。但是Zillow网上明确地说明，Z评估基于房屋“当下”的价值。[注]鉴于Z评估总是每周升级三次，在你评估房屋到卖出房屋中间这段时间内，你所在区域（或者其他因素）房屋销售额有可能会改变一座房屋的Z评估。在此有两个启示——首先，在你使用数据时，弄清楚你看到的数据是不是预测值。其次，确保将预测偏差考虑

在内——无论它已经计算在内了，还是你正因其对预测做出调整。

最后，想想Z评估是通过什么得出的。是通过一个专有公式得出的，这意味着你无法了解其中涉及的所有因素。至少它部分依赖于自陈式数据，我们知道（在一些情况下）它们不是百分之百可靠的。而可用数据的数量可能随着地区不同而不同，或者同一地区内，每周都发生变化。如果你没有数据，那你就需要用上你手中的样本。如果某一地区有较多的交易量，这会为Zillow提供更多数据，然后Z评估可能会更精确些。

或者，我们觉得Zillow会说——“贼精确”（Zaccurate）。

---

1. We say “nearly” any home because Zillow doesn’t appear to track every home in the U.S., based on data we compared from the American Housing Survey from the U.S. Census Bureau. (The United States Census Bureau, American Housing Survey (AHS), last revised May 14, 2015.) That said, the differences we found could be due to time (the data was collected approximately two years apart), or due to definitions in terms of what constitutes a “housing unit” (for the bureau) versus a “home” (for Zillow). The data we looked at includes: Table C-01-AH,American Housing Survey, accessed August 6, 2015, http://www.census.gov/programs-surveys/ahs/data/2013/ national- summary- report- and- tables--- ahs- 2013.html. This number is itself based on a sample of 70 million housing units.We also looked at “Zestimate,” Zillow website, accessed August 6, 2015, Zestimate_Accuracy_2015_03_31, http://www.zillow.com/zestimate/#what.

2. Zillow offers quite a bit of information about its Zestimates— including a downloadable Microsoft Excel spreadsheet that shows their accuracy (we found the link here: http://www.zillow.com/zestimate/) and a website dedicated to data(“Zillow Real Estate Research,” Zillow website, accessed August 6, 2015, http://www.zillow.com/research/data/).

3. Kenneth R. Harney, “Inaccurate Zillow ‘Zestimates’ a Source of Conflict Over Home Prices,” Los Angeles Times website, February 8, 2015, http://www.latimes.com/business/realestate/la-f- harney- 20150208- story.html.

4. Keep in mind that real estate agents may have less to lose than you do when it comes to reducing the price of your house. For example, if your house sells for$300,000, your agent might get $5,250 (a 7 percent listing fee = 3.5 percent for the buyer’s agent, and 3.5 percent for the seller’s agent; of that 3.5 percent, half may go to the agent, and half to the agency, so each agent ends up with 1.75 percent). If you lower the price of your house and it sells for $250,000, your

agent only loses $875 (1.75 percent of $50,000) but you've lost a lot more.

5. David Howell, "How Accurate Is Zillow's Zestimate? Not Very, Says One Washington-Area Agent," Washington Post website, June 10, 2014, http://www.washingtonpost.com/blogs/where-we-live/wp/2014/06/10/ how- accurate-is- zillows- zestimate- not- very- says- one- washington- area- agent/.

6. The study found that initial Zestimates "were within 5 percent of the ultimate sale price 46 percent of the time," while real estate agents' initial list prices were within 5 percent "76 percent of the time." Stan Humphries, "How Accurate Is the Zestimate? Zillow Says the Tool Is Helpful When Used the Right Way,"Washington Post website, June 10, 2014, http://www.washingtonpost.com/blogs/where-we-live/wp/2014/06/10/ how- accurate-is- the- zestimate- zillow- says- the- tool-is- helpful- when- used- the- right- way/.

7. Note that we looked at data from a specific day, and that the values may have changed between then and whenever you're reading this.

8. Zillow says stars are "tied to" the median error rate. In the data we reviewed, a median error rate of 5.3 percent to 7.4 percent got four stars, 7.6 percent to 8.9 percent got three stars, 9.1 percent to 11.8 percent got two stars; and areas with one star didn't have a median error rate given.

9. Including a four- star rating and the lowest median error rate for top metro areas,which was 5.0 percent when we reviewed it on September 1, 2015 (last updated August 26, 2015). "Zestimate," Zillow website.

10. "Zestimate," Zillow website. Although we should note that Zillow does offer a Zestimate forecast for some properties (http://www.zillow.com/blog/ zestimate- forecast- 151664/), and it has a Zillow Home Value Index (http://www.zillow.com/ home- values/) that offers predictions about the housing market.

如何成为一个成熟的数据接收者

一如往常，在本章的结尾，我们列出5件想成为一个成熟的数据接收者现在就可以做的事。由于这是本书最后一章了，所以我们从整体着眼。下面是我们希望你们在日常生活中遇到数据时需要铭记的5个要点。

1.当你看到、听到数据时，要学会识别数据。一篇报纸上的文章、广播故事、来自供应商的电子邮件、孩子的成绩单、下周的销售额预测、地图等全部都是数据。不论你住在哪里，在做何事，你每天都可能被数据所围绕。

2.确保事实正确。许多数据问题只是一个错误导致的。或许电子表格中的公式用错了，或者关键数值的小数点放错了位置。也许一位博主无意间曲解了一项最新的科学研究。你首先应该做的一步是核实你看到的数据是否正确。

3.了解数据来源，以及展示这些数据的人是谁。在有些情况下，个人或者组织可能会有周密的计划，这意味着他们可能会调整数据（你也可以称之为筛选数据），使其符合他们要传达的信息。毕竟，你通常不会听到民主党总统候选人赞扬共和党人领导的倡议活动（反之亦然）。甚至在没有明显目的的情况下，你接收到的数据也有出入，以某种方式被搜集并传播——能够影响最终结果的所有因素近在眼前。

4.留意明显的数据陷阱。你很有可能会打开家乡的报纸（或者浏览网站），然后发现一则新闻，新闻中的数据仅仅证明了关联性，而字里行间却在暗示因果性。以我们的经验来看，正常情况下，你最有可能遇到的一些其他因素是什么？样本数量小、结果存在非显著性差异（或者

存在显著性差异，不过带来的效应很小）、欺骗性平均值以及包括信息可视化在内的误导性可视图形。

5.要知道，正确解读数据能够帮助你做出更好的决定。最后，这才是本书的初衷——分析数据以解答关乎自身的问题。下一季度你公司的销售额将会有多高？关于致癌食品的最新研究是你应该担心的事情吗？买一栋避暑别墅，什么样的价格才算合理？发现对你的生活影响最大的问题，然后看看你会如何运用本书里的方法解答问题。

尾声

到此就结束了。

其实并没有。

因为当你第二天早上醒来的时候，你会被更多的数据狂轰滥炸。更多的抽样及筛选数据的情况也会发生，更多的人会混淆关联性与因果性，会有更多的网站、博客、新闻广播不停给你灌输你该或者不该做什么，从而让自己寿命更长、头脑更灵、活得更好。

但愿，读完本书，你能够运用所学到的知识，看到隐藏在你每天都会用到的所有“小数据”背后的错误信息，并清楚地知道如何应对。

词汇表

注意：这些是一些常见术语的一般含义。

合计数据（aggregated data）——单个数据点汇集而成的集合（例如将每个州的总选票数整合起来，以此判断谁会赢得总统选举团的选票）。

平均值（average）——概括性统计的一种（通常是平均数、众数或者中位数），以单一指标描述数据。

大数据（big data）——是指由于数量过于庞大，不借助精密仪器或者高级计算机就无法处理的数据。

双变量关系（bivariate relationship）——说得直白点，即两个（双）变量之间存在的关系（例如，房价与卫生间的数量有关）。

黑天鹅事件（black swan event）——可能性非常小的事件，而一旦发生其影响巨大。

因果性（causation）——在该关系中，一种因素会引起另一种因素发生。

筛选（cherry-picking）——从数据中选择传闻的例子来说明自己的观点，同时忽略其他可能与观点相悖的数据点。

置信区间（confidence interval）——衡量关于结论数据确定程度的方式，往往表现为一个范围值，置信区间体现了一个可能看到估计结果的范围（假设你有一个随机的、有代表性的样本）。

置信水平（confidence level）——我们用该术语判定对正确衡量数据的自信度。

实证性偏见（confirmation bias）——用巩固自己预期的方式解释数据的倾向。

关联性（correlation）——两个变量之间的数据关系，通常定义为正相关（朝同一方向发展）或者负相关（彼此朝着相反方向发展）。

数据（data）——信息或者事实。

依赖性（dependence）——一种变量直接由一种变量决定。

确定性预测（deterministic forecast）——能确定地得出精确结果的预测。（例如，明天上午9点我家会下雨。）

经济影响（economic impact）——事情会将会耗费多少时间、金钱、健康等资源。

估算值（estimate）——从一个简单的样本数据就能推断出整体的数据。

小数据（every data）——我们用该词来描述日常数据。

外延有效性（external validity）——把从样品中得出的结论加以推广，从而得出对整个数据总体有意义的结论。

假阳性（false positive）——数据预测预测出虚假结果的情况。

预测（forecast）——有关未来的说法，尽管对特定人群来讲，预测与预知含义可能不同，但除非特殊说明，我们一般将二者作为同义词使用。

预测偏差（forecast bias）——该术语用来描述一贯偏高（正偏差）或者偏低（负偏差）的预测误差。

推测（inference）——得出相关数据结论的过程。

效应尺度（magnitude）——从本质上说，指效应的大小。

误差范围（margin of error）——衡量数据不确定性的方式。

平均数（mean）——当你说“平均值”时，多数人想到的词（你将所有数值求和，然后除以数据个数就可以得到平均数了）。

中位数（median）——位于排序数据集最中间的数值。

曲解（misrepresentation）——数据以不正确或容易引发误解的方式呈现。

众数（mode）——在数据中出现最频繁的数据点。

观测（observation）——观察诸如人、价格或者某天的数据单元。

比率（odds）——在统计学中，某事发生的比率是指某一结果发生与不发生的可能性的比例（例如一匹马赢得一场比赛的概率可能是1/3，意思是这匹马每三场比赛会赢一场，在博彩术语中，比率往往是相反的，同一匹马获胜的比率是2∶1，意思是输掉比赛的可能性为2/3）。

遗漏变量（omitted variable）——在一种关系中起作用，但是可能被忽视或者被排除掉的变量。遗漏变量是关联性不等于因果性的主要原因之一。

离群值（outlier）——不合群的特殊观察结果，它可能远超（或者低）于其他所有数据，或者它只是没有遵循事情的一般模式。

P值黑客（P-hacking）——名字来源于P值，P值黑客是指反复分析数据，以期让不存在显著性差异的结果变得存在显著性差异。

P值（P-value）——衡量显著性差异的方式，P值越低，你看到的结果源于偶然的可能性越低。

数据总体（population）——你想要研究并得出相关结论的数据或者观察结果的完整集合，统计学家在研究中很少能掌握全部数据，但掌握一个小型的、经过精心挑选的数据组是可能的。（如，100名美国议员的投票习惯。）

预知（prediction）——参见“预测”。

预知偏差（prediction error）——衡量未来不确定性的方式，主要通过对比事情发生时的实际结果与预测结果得出。

预测区间（prediction interval）——下一个数据点可能出现的范围。

可能性预测（probabilistic forecast）——判定结果发生概率的预测（例如，明天有雷雨的概率为30%）。

概率（probability）——发生结果的可能性（往往用百分数、分数或者小数表示）。

指标（proxy）——与另一个难以衡量的因素紧密关联的因素（例如，智商是先天能力的一个指标）。

随机（random）——所观察到的东西出于偶然，而不是一个可以

总结出来的过程或事件。

风险（risk）——该词的含义因人而异，一般来讲，风险不仅同事情发生的概率有关，也与其后果有关。

样本（sample）——全部数据中的一部分（例如，“挑战者号”存在O型环故障的发射任务）。

样本选择（sample selection）——当选择样本的方式直接与研究中的结果相关时所出现的潜在统计问题，另外有时用于描述从全部数据中选取样本的过程。

样本误差（sampling error）——由于不知道样本是否代表全部数据真实数值而导致的不确定性。

选择偏差（selection bias）——由选择参与的人组成样本时会出现的潜在问题，该问题可能会影响结果。

虚假关联（spurious correlation）——没有实际的或者经济意义的数据关系，或者是受遗漏变量影响的关系（例如，谋杀率与冰淇淋销量之间的关系）。

统计（statistic）——以数字的方式描述数据某一方面（如平均数、中位数、众数等）。

统计影响（statistical impact）——产生了显著性差异，但未确定其效应的大小。

显著性差异（statistical significance）——一种以概率为基础的方式，用以断定观察影响是以数据的形式呈现出来，还是仅仅随机出现。

概括统计（summary statistic）——该概念提供数据一个或多个方面的信息。平均数和数据合计就是概括统计的两个实例。

加权平均值（weighted average）——（根据数值的相对重要性）为数值加上权重之后计算得出的平均数。

致谢

本书得以完成，离不开一路上相信我们并支持我们的许许多多的人。

本书从草稿到完成，得到了许多同事、朋友无价的指教、细致的评点，特将姓名罗列如下：尼尔·博格、史蒂夫·布朗纳、杰弗里·布朗、科特尼·科勒、德瑞克·丹尼尔、罗恩·弗里德曼、卡拉·格尔斯基、鲍勃·霍夫曼、艾米丽·欧斯特、J.J.普雷斯科特、凯文·斯默克勒、理查德·施特劳斯、泰勒·维根、贾斯汀·沃尔夫斯。还要对埃奇沃思午餐小组的各位成员致以热烈的感谢，因为他们每天都在谈论数据。

对于本书的另一批主要贡献者包括吉尔夫·巴尼斯、比尔·福特、瑞贝卡·格林美尔、凯迪·郝尔、乔丹·海格依、弗兰克·霍洛维奇、华许峰（比尔）、约翰·吉罗迪、金俊洙、金米亚、马尔蒂·马丁、约翰·麦基翁、凯尔·墨里西、皮特·莫非、卡西·雷恩哈德、切尔伊·卢登、乔尔·卢登、罗西·松西尔莫、本·怀特纳。

统计学和经济计量学不是靠一个人凭空想就能学会的，约翰多年来受到几位杰出的经济计量学教授的点拨和指导，这些教授是梁晓晖、杰瑞·豪斯曼、史蒂夫·皮斯切克、约书亚·安格里斯特。

我们非常感谢我们的代理人特里斯·考伯恩，以及在Bibliomotion出版社的整个团队，成员分别是艾瑞卡、吉尔·F、吉尔·S、艾丽西亚、艾瑞、席福恩、苏珊、苏。

不过，我们把最由衷的感谢留到了最后。因为这本书的撰写到出版这一整个过程，只有我们的朋友以及家庭成员一直陪伴着我们。嗯，在

大多数情况下他们会耐心地看着我们耗费无数时间校对书稿。我们希望这样的等待是值得的。

迈克尤其想谢谢一直支持着他的父母，以及时常会提出很有道理观点的哥哥丹尼尔，还有他的孩子扎克和本（看，孩子们，你们的名字印在书里了）。当然，还要谢谢他的妻子马尔拉一直以来的宽容相待。

约翰要感谢他的父亲以及姐夫吉姆和杰克，他们对约翰写作此书一直很感兴趣，表现出十足的热情。约翰还要感谢他的弟弟凯文和史蒂夫（作家协会会员）的鼎力支持。约翰十分幸运，有着一帮全身心支持他的朋友，包括吉博力·艾门、帕特里克·拜恩、杰西·大卫、德博拉·福斯特、查克·菲尔德、莱拉·海德尔、克里斯·约翰逊、迈克·凯菲特、帕克尔·诺曼、纳珊·伍德、迈克·维尔、乔治·科伦科、马特·米尔纳、史蒂夫、舒尔曼。这些人会定期聚在一起讨论这本书。感谢他的孩子科特尼和马修，他们非常想要了解数据，因为约翰认为这很重要（还要感谢他们真的成功地把这本书预售给了朋友们）。最后，当然也是十分重要的，就是感谢约翰的妻子克里斯蒂，对于约翰以及他的疯狂想法，她自始至终都很支持。